

Recognizing hollow strengths in research communities

A review of physics in the United Kingdom highlights the usefulness of external scrutiny and the challenge of getting a part of the community to set itself new standards. Who will take the lead?

How best to assess a country's scientific research? Bibliographic indicators are important, but there is no substitute for getting authoritative and, hopefully, wise people to see for themselves and express their views. In fact, the 11 eminent physicists who have just presented a report, *International Perceptions of UK Research in Physics and Astronomy*, had more than just impressions to go on. The Institute of Physics supplied them with data from the UK university research assessment exercise, levels of funding with international comparisons, and demographics. They conducted a survey of 150 physicists outside the United Kingdom. They spent about a week visiting half a dozen labs in the United Kingdom and reviewing the evidence. The results can be found at www.iop.org/Policy/Intrev.html.

As the panel emphasize, their visits were brief. "The word 'perceptions' is apt," they acknowledge. Nevertheless, they doubt that a longer visit would have changed their perceptions significantly. The problems they point to, with the additional benefit of their perspective, are all too glaring.

The panel consisted of six physicists from the United States and five from continental Europe. By all accounts, the lack of UK participants was a strength for the obvious reason: there was no tendency to pull any punches. The presence in the report of some polite but pointed criticisms and concerns adds force to the panel's more positive overview: by and large, astronomy and physics in the United Kingdom are holding their own at the international cutting edge. But there is a clear warning to be gathered from the report: in some critical aspects, that scientific strength rings hollow.

Small physics

According to the wise men (unsurprisingly for physics, there were no women on the panel), worries are most acute in 'small physics', and in three areas in particular: economically strategic physics, the state of infrastructure, and fundamental atomic and molecular physics.

No country interested in drawing economic strength from science can afford to ignore research into the properties and processing of materials at the nanometre scale — as the United States, Germany, France and, for many years now, Japan have perceived. Here, both in science and in its collaboration with engineering, the United Kingdom is seen by the panel as weak. That testifies to a lack of industrial strength, too, and an apparent failure of the government's long-established foresight and related initiatives to stimulate a key area of development. Perhaps the United Kingdom can learn from the United States' new drive in this direction, which is drawing on the strengths and motivation of both defence and civil agencies.

The report also points to dismal infrastructure, highlighting limitations of another notable government development, the Joint Infrastructure Fund, which had originally been intended to renew equipment and labs for current research programmes. Assertions are now emerging that it has instead been devoted too much to new research projects. Moreover, infrastructure for small physics appears

to have failed to attract funding because of the need to spend much larger sums on other sciences. But heads of research funding agencies also report a wealth of excellence in the applications for infrastructure that the money could not accommodate. Clearly, the fundamental message is that the UK infrastructure needs much more by way of renovation if the United Kingdom is to remain internationally impressive.

Derivative work

That said, infrastructure is only worth investing in if the science it will support is world-class. But what if an external review says the following: "In comparison with earlier times when the UK was recognised as a leader and innovative, the work now is regarded as largely derivative"? Officials and competitors for funds could seize on this as an excuse to stop funding in that area. But this comment refers to atomic and molecular physics. This science probes the fundamentals of quantum theory and is creating exciting new states of matter and powerful new applications. There are pockets of forward-looking excellence in the United Kingdom, but one is left in no doubt that, in the eyes of this panel, too many atomic and molecular physicists have allowed the rest of the world to leave them stranded up a backwater.

Who is to blame, and what is to be done? Over recent years the funding agency responsible for that area — the Engineering and Physical Sciences Research Council — has blown lukewarm and, more recently, warm over fundamental physics. But the inward-looking character of UK atomic and molecular physics, as perceived by these outsiders and by some insiders, too, has surely become entrenched over much longer timescales. In fairness, this particular community has yet to respond to the report. But it is not easy in such a situation for a leader to step forward with a rallying cry. There needs to be a combined pressure of top-down reluctance to fund second-rate research as judged by international standards, and bottom-up pressure through discussion, leading to community support for sharper funding criteria. This will take years unless leading groups can be supported strongly at the expense of others.

But there are broader lessons, too. Countries who haven't tried similar exercises should consider doing so in the light of this example. Meanwhile, in the United Kingdom, for the institutions that steered this exercise — the Royal Astronomical Society, the Institute of Physics and the relevant funding agencies — the report is well timed: the British government is reaching a critical moment in a review of government spending. The government can be pointed to an independent assessment of the country's great potential in physics and astronomy, as well as critical shortages in funding and, yet again, to the need for more support for younger scientists. But they and UK industry need to do a better job of convincing government that physics is worth supporting for its own fundamental and exploitable aspects — provided everyone can see that truly world-class science will be the outcome. ■





Frog alert
Initiative will investigate decline in amphibians
p495



Opening port
Brest to be home for French research fleet
p498



Clean sweep
Dredger captures chimney from marine volcano
p499



A moving tale
Science academies to set up office in Trieste
p501

Researchers take US government to court over threat to turtles...

Natasha Loder

US marine biologists have drawn a line in the sand. In the face of rapidly dwindling numbers of Pacific leatherback turtles, they have decided that "enough is enough" and are suing the US government for failing to protect the species.

Leatherbacks are endangered throughout the world. But the numbers of the genetically distinct Pacific leatherback turtle are declining precipitously, probably because of their accidental capture by longline fishing. They become entangled in lines or hooks, and some either drown or get picked off by sharks, while many of those released by fisherman later die as a result of their capture.

Although the National Marine Fisheries Service (NMFS), the US government agency responsible for conserving and managing marine resources, recognizes Hawaiian longline fishing as the main threat to the recovery of the leatherbacks under their control, it is now permitting more accidental captures.

Last year, the service authorized an unprecedented take of 244 leatherback turtles, based on a claimed population figure of 85,000 leatherbacks. Critics dispute this population total, as well as the assumption that only a handful of those caught will die. Scientists from the Center for Marine Conservation (CMC) in Washington are now accusing the NMFS of failing to perform its mandatory duties under the US Endangered Species Act.

Pamela Plotkin, a senior conservation scientist at the CMC and the lead scientist in the legal action, says her unit had been in dispute with the NMFS over this issue for several years when it was approached by the Earth Justice Legal Defense Fund, based in Honolulu.

The non-profit law firm works in the area of environmental justice and was looking for plaintiffs to challenge the many unmanaged Hawaiian fisheries. Plotkin agrees that the action is unusual for a group of scientists, but adds: "as you work on different organisms you become passionate and committed and feel it is part of your job to speak out."

Plotkin has already testified that government figures are incorrect, and several prominent scientists have written to the judge warning him that the number of nest-



No place to hide: accidental captures in longline fishing have been a major factor in driving the leatherback turtle close to extinction.

ing turtles could be as low as 3,700, and that they are in imminent danger of extinction.

In October last year, the judge ruled that the NMFS had violated the National Environmental Policy Act by allowing a Hawaii-based longline fishery to operate without first completing an environmental impact study. He then closed thousands of square miles of the Pacific Ocean to longline fishing until the study had been completed.

Scientists from Drexel University in Philadelphia and the CMC have now produced a paper (see pages 529–30) concluding that the leatherback is on the verge of extinction in the Pacific. Their study population in

Costa Rica has fallen from 1,367 nesting females to 117 in ten years. Estimates of annual mortality for this and an important Mexican population are 35 and 20 per cent.

Jim Lecky, the assistant regional administrator for protected resources at the NMFS, says they expect to learn a lot from the lawsuit, but warns: "This is a huge international problem and we could shut down Hawaiian longline fishing tomorrow, and it would not make a dent in this." Lecky also argues that it would be unfair to put Hawaiian fishermen out of business while other countries would be able to continue fishing. A final decision is due in the next few months. ■

...as amphibians come under study

Rex Dalton, San Diego

A disturbing trend of population declines and birth deformities in amphibians has prompted a US federal agency to launch an ambitious research programme.

The national Amphibian Research and Monitoring Initiative (ARMI) has been set up by the US Department of the Interior with \$5 million to study why frogs, salamanders and toads are in such trouble. An additional \$2 million is being sought for next year's budget.

Initially, ARMI will examine amphibians on the nation's extensive federal lands, but officials hope to expand the programme to

other habitats. Coordinated by the US Geological Survey (USGS), the scheme was created after scientists captured the attention of Interior Secretary Bruce Babbitt with reports of amphibian distress.

"This is now a front-burner issue," says James Hanken, herpetology curator at Harvard University's Museum of Comparative Zoology. "This is not just a plan; things are happening."

Hanken was among a group of scientists who met early last month in West Virginia to design ARMI. Officials from various federal agencies, including the National Science Foundation, the National Institutes



In the ARMI now: the yellow-legged frog is one of many amphibians under threat.

of Health and the Environmental Protection Agency, also took part.

Under ARMI, the nation has been divided into seven regions of particular amphibian characteristics, says Daniel James, a USGS wildlife biologist who is coordinating the programme. Government agencies had strong herpetological staff in four regions, but not in the Mississippi River Valley or the Northeast, so the department is seeking new scientists to augment its resources.

Initially, researchers will survey many sites in large sections of National Parks, National Wildlife Refuges and Bureau of Land Management areas. They will then focus on a few species at a limited number of locations in an attempt to determine what ecological factors are contributing to the declines and deformities.

The ARMI effort is taking place at a time when scientists around the world are seeking better understanding of the broad declines taking place in global amphibian populations. Pollutants, increased amphibian susceptibility to disease, and decreases in the ozone layer are all being considered.

An international group of scientists, the Declining Amphibian Populations Task Force, has played a major role in spurring amphibian research. ARMI is to become the permanent home of an amphibian atlas that is now being developed, officials said.

In addition, federal officials have hired an amphibian pathologist at the National Wildlife Health Center in Madison, Wisconsin, to focus on the role of disease. ARMI also complements an on-going three-year amphibian study by the NSF, which is already providing \$3 million to 13 universities to study the causes of species declines.

Green light for plans to sell off US helium reserve

Colin Macilwain, Washington

The US government's plans to privatize the nation's helium reserve have been backed by a panel at the National Research Council (NRC). Some physicists had attacked the proposals because of the limited supply of this inert gas. But the NRC panel found that privatizing the reserve would not harm long-term supplies to researchers and other users.

The national reserve holds 30 billion cubic feet of helium — enough for ten years at current usage rates. The gas is extracted from natural gas supplies in the United States, and the physicists were worried that there would be no other source when these supplies are exhausted. Helium is used mainly for industrial cleaning and welding, and for cryogenic research.

A statement issued in 1995 by the American Physical Society (APS) expressed "profound concern" about the sale of the reserve, noting that the supply of helium has "severe natural limits". Despite this, legislation was passed in 1996 to sell off the reserve, subject to a report from the NRC and subsequent consultations with helium users.

The panel's report dismissed the physicists' concerns. The panel found that "the Helium Privatization Act of 1996 will not have a substantial impact on helium users". This clears the way for the sale of the reserve.

APS officials endorsed the composition of the NRC panel, which was chaired by Ray Beebe, a consultant engineer, and John Reppy, a physicist at Cornell University, and are expected to accept its findings.

According to Beebe, the panel looked at

other materials whose sole source is as a by-product of another commodity — including several rare metals that are only extracted as offshoots of copper mining — and concluded that such supply restrictions were unlikely to lead to a helium shortage.

The panel also found that most of the current demand for helium was being met by sources independent of the national reserve. To safeguard long-term supplies of the gas, the NRC panel recommends more research into natural gas fields with high helium content, and on helium conservation and recycling.

Asked if physicists were wrong to say that the privatization would endanger their helium supplies, Beebe said: "People who are focused on their own work will naturally be very protective of it."

♦ <http://books.nap.edu/catalog/9860.html>



UK labs impeded by old equipment

Natasha Loder, London

More than half of the university laboratories in the United Kingdom are using outdated equipment, according to a report from Manchester University's Policy Research in Engineering, Science and Technology Institute. It says that at least £600 million (US\$885 million) is needed to bring essential research facilities up to date.

The study looked at four research-intensive universities and analysed applications for recent bids to a special £750 million research infrastructure fund. It found that that two-thirds of departments had critical experiments they were unable to perform because of a lack of equipment.

Several department heads said their research strategies were driven by existing or

prospective non-availability of equipment. The study also found that the fund, intended to remedy a two-decade hole in underfunding of research infrastructure, was hugely oversubscribed, falling far short of what is needed and leaving many applications rated 'excellent' unfunded.

The work was commissioned by the Committee of Vice Chancellors and Principals (CVCP), which represents the heads of UK universities. It was intended to focus minds at the Treasury, which is reviewing all areas of government spending before allocating funding for the next three years. Diana Warwick, the CVCP chief executive, describes the findings as "important", and emphasizes the "urgent need to address basic funding priorities for UK universities".

Clinical trials end at gene-therapy institute...

Paul Smaglik, Washington

Reflecting suggestions that it may have been over-zealous in promoting the human applications of gene-therapy research, the University of Pennsylvania announced last week that its Institute for Human Gene Therapy (IHGT), which was designed to embrace a 'bench-to-bedside' approach, will no longer conduct clinical trials.

The decision follows investigations into the death of an 18-year-old man, Jesse Gelsinger, in one such trial at the institute last autumn (see *Nature* **401**, 517–518; 1999). Gelsinger's death and the ensuing fallout have already put the field under greater scrutiny, and could lead to regulations that would make conducting any clinical trial more difficult and more expensive.

Part of the fallout has been the suspension by the US Food and Drug Administration (FDA) of all five of the IHGT's trials, amid revelations that many other US gene-therapy investigators failed to report adverse events to the US National Institutes of Health.

Last week, the Department of Health and Human Services (DHHS) called for better training of clinical investigators, stricter guidelines on obtaining informed consent for experiments, and more aggressive monitoring of clinical trials. These recommendations reflect criticisms made by the FDA when it suspended the IHGT's clinical trials (see *Nature* **403**, 354; 2000).

Senator William Frist (Republican, Tennessee), chair of the Senate's Subcommittee on Health, Education, Labor and Pensions, noted during a subcommittee hearing last week that these recommendations resemble those made in 1998, which a recent progress report says the department has largely failed to meet.

Frist added that the new DHHS guidelines ignore many other recommendations made by the department's Office of Inspector General, particularly on improving the review boards that monitor clinical trials for research institutions.

The DHHS has already said that it intends to introduce laws under which investigators could be fined up to \$250,000, and institutions up to \$1 million, for failing to meet the stricter standards. But Frist suggested that instead further laws may be needed to make researchers follow more stringent reporting and monitoring requirements.

An external report commissioned by the university on Gelsinger's death, and published last week, concluded that tougher requirements would be costly. "Since the university will be judged by the same standards as industry, it will now be necessary to invest considerably more resources for compliance than has historically been necessary for independent university-initiated research," it says.



Wilson: will remain at the institute, but observers believe he was 'overloaded'.

The University of Pennsylvania (Penn) invested heavily in the IHGT, which has some 250 employees and an annual budget of \$25 million. Judith Rodin, the university's president, says this investment was not enough. But rather than build up the institute further with its own funds, Penn will look outside for help.

The university is to hire a contract research organization to handle the overseeing of all clinical trials — not only those involving gene therapy — that its Office of Regulatory Affairs considers risky or complicated, and which are not sponsored by pharmaceutical companies.

At the same time, the IHGT will be made smaller, both in terms of budget and staff, although Rodin declines to provide solid numbers. The institute's director, James Wilson, will stay on, but will not be allowed to participate in any clinical trials.

Rodin and the report's lead author, William Danforth, chancellor emeritus of

Washington University in St Louis, are saying publicly what many gene-therapy researchers have said privately since Gelsinger's death — that Wilson extended himself too far. Besides his role as institute director, Wilson was involved in several clinical trials, as well as in basic research. "We think Dr Wilson was overloaded," says Rodin.

Penn's relationship with the biotech company that produced vectors for the institute is also uncertain. Genovo, Inc., based in Sharon Hill, Pennsylvania, was founded by Wilson, who has financial interests in it. Its contract with the university ends in June. Rodin declined to say whether it would be renewed.

While Wilson's holdings in Genovo suggested a potential conflict of interest, Danforth says his committee found no evidence of wrongdoing. However, the committee's report states that the public's perception of conflict of interest can be nearly as dangerous as the real thing.

The university may apply to the FDA to run the IHGT's gene-therapy trials through other university programmes. The report questions whether the university should dedicate an entire institute to gene therapy.

"Devoting an institute to human gene therapy suggested a little more confidence that gene therapy could work than might have been justified at the time," says Danforth.

He agrees with the FDA's earlier assessment that Penn made mistakes during the trial, including the monitoring and reporting of adverse effects, but wasn't sure of their impact. "Did those things criticized by the FDA lead to the death of Jesse Gelsinger? I think there is no evidence of that," he says. ■

...as Harvard keeps its ethics guidelines

Steve Nadis, Boston

The Dean of Harvard Medical School (HMS), Joseph Martin, announced last week that the school will not relax its conflict-of-interest guidelines, as had been recommended by a senior faculty committee in a report last autumn (see *Nature* **403**, 818; 2000).

The policy, said Martin, would stay as it is except for new safeguards imposed to protect students and other trainees from "potential conflicts created by their mentor's financial interests".

Dennis Kasper, a member of the faculty committee and the executive dean for academic programmes at HMS, notes that

the committee's decision was based on "an evolving faculty point of view and was not influenced by outside forces", such as recent critical editorials in medical journals.

But he did say that Martin's participation on a panel that investigated the recent death of a patient in gene-therapy trials at the University of Pennsylvania (see main story) "supported the direction he'd been leaning in all along".

Some medical school researchers are critical of the decision. Cell biologist and haematologist Thomas Stossel, for example, argues that the current policies inhibit the development of

valuable drugs. "When researchers discover things that might lead to useful products, I think they are ethically obliged to steer it to the public by going through industry," he says.

Both Martin and Kasper leave open the possibility that Harvard's restrictive guidelines could be changed in the future to bring them more in line with those of other institutions. But they stress that this should only occur after a national forum is held by the leading research universities, government, and industry in the hope of establishing national standards. "All of us should stand by the same ethical rules," says Kasper. ■

France plans marine research port at Brest

Declan Butler, Paris

Nazi Germany recognized the excellent strategic position and geography of the French port Brest, and installed its U-boat base there during the Second World War. Now France plans to spend FF39 million (\$5.4 million) building a scientific port there to support its extensive research fleet and oceanographic campaigns.

The French marine research agency, IFREMER, has the second largest research fleet in the world after the United States. With seven large research ships, two manned submarines that can dive to 6,000 metres, and a series of robotic and other seagoing facilities, the fleet is comparable to that of the United Kingdom.

Around half of the agency's 1,700-strong workforce is already based in or around Brest, which is also home to marine laboratories from several universities and the national basic research agency CNRS. The other large concentration is in Marseille, for Mediterranean-related research. Brest is also home to the French Institute for Polar Research and Technology.

But its working conditions are less than ideal, explains Joël Quérellou, director of the IFREMER centre in the city. Research vessels share the commercial port, and specialized support and research facilities are lacking. IFREMER's own laboratories are spread out over four sites.



Splashing out: *Atalante* is part of the IFREMER fleet that may move to the new port at Brest.

The new plan, says Quérellou, who is the main architect of the project, will concentrate 200 support staff and facilities in a dedicated scientific port alongside the commercial port.

The infrastructure available at Brest's vast military port is better than that available at the trading port. But the option of locating the scientific port there has been discounted, mainly because security considerations could delay ocean research campaigns involving foreign scientists.

The money for the project is coming from regional and state funding, and has not been difficult to muster. The military arsenal is in decline, with job cuts of 350 this year reducing the workforce to 4,000, and the government is keen to boost alternatives for growth in the region.

This week, the Océanopolis marine theme park reopens in the area, for example, after a FF250 million refurbishment. Its vast aquariums and displays, the outcome of a joint effort between the local scientific community and the centre, are expected to attract 600,000 visitors annually.

Quérellou says that in the past these local factors have been imperative, but he now intends to explore Brest's role in Europe's overall marine research. Not before time, says Pierre Papon, a former director of IFREMER. He argues that national initiatives can no longer afford to be pursued in isolation from their European context.

Although France, Britain and Germany have agreed to cooperate in the use of the 40 or so existing research vessels (see *Nature* 379, 576; 1996), and some ships have been built jointly—the Franco-Spanish *Thalassa*, and a smaller Franco-Italian ship, *L'Europe*, for example—there is a need for greater cooperation, such as in the planning of future needs for vessels.

A proposal to create a European maritime research agency (see *Nature* 392, 323; 1998) now seems unlikely to be realized, a lighter coordination being preferred by many European countries. This idea will be on the agenda of a European Union-sponsored conference, Ocean 2000, to be held in Hamburg, Germany in August. ■

G. VINCENT/IFREMER

Transmutation of nuclear waste branded 'Trojan horse'

Colin Macilwain, Washington

A technique for disposing of radioactive waste, under study by many nuclear powers, was last week branded "a Trojan horse for perpetrating nuclear power and hence the generation of more and more radioactive waste" by a respected environmental group based in Washington.

The technique, known as transmutation, is sometimes touted as a safe and cost-effective way of disposing of nuclear waste. The governments of both France and the United States, for example, are currently weighing up proposals to expand their research into the field.

Transmutation would use neutrons from an accelerator or a nuclear reactor to break down radionuclides with long half-lives. Advocates say it could enable nuclear waste to become safe after a relatively short period of storage.

A team of US scientists at the Los Alamos National Laboratory in New Mexico has just prepared a 'technology roadmap' for the development of transmutation techniques.

This is now being considered by the US Department of Energy.

Meanwhile the French Atomic Energy Commission is refurbishing its PHENIX fast reactor at Marcoule, near Avignon, to enable it to conduct transmutation research before the reactor closes in 2004.

The commission is waiting for formal safety approval from the Direction de Sûreté des Installations Nucléaires before it can restart the reactor and allow the research programme to proceed.

But the Washington-based Institute for Energy and Environmental Research (IEER), led by physicist Arjun Makhijani, argues that the technology can never be made to work, as it cannot break down some of the most troublesome radionuclides, including strontium-90 and caesium-137. The IEER also says that transmutation would create new types of waste, "along with new proliferation risks and high costs".

Greg Van Tuyle, leader of the Accelerator Transmutation of Waste (ATW) group at Los Alamos, counters that the technology is

"threatening to some groups primarily because it could help to restore the viability of nuclear energy at a time when it is becoming essential to addressing global climate change". He adds that transmutation "is the responsible approach" and would leave waste "in a form far less vulnerable to environmental release".

The IEER's scepticism about transmutation technology is widely shared, and the US Department of Energy has not allocated any money for its study in recent years. But Senator Pete Domenici (Republican, New Mexico), who chairs the appropriations subcommittee that funds the department, added \$9 million in the current fiscal year to pay the team at Los Alamos to prepare its transmutation roadmap.

A congressional official, who asked not to be identified, claims that the IEER report contains technical errors and "presumes the results of research that hasn't been done yet". The official says that Domenici "wants to avoid shutting down any possibility" for the safe disposal of nuclear waste. ■

Germany edges towards stem-cell accord

Quirin Schiermeier, Munich

Political agreement over the use of human embryos in research remains elusive in Germany. But, judging by a ground-breaking three-day symposium held in Berlin last week, the gap between the extreme positions is narrowing.

Organized by Germany's federal health minister, Andrea Fischer, the meeting addressed the scientific, medical, ethical and legal aspects of reproductive medicine. In particular, the aim was to lay the groundwork for a possible revision of Germany's embryo research laws — currently the strictest in Europe.

The meeting was intended to complement the work of an all-party parliamentary commission, set up earlier this year to identify the areas where scientific advances have exposed a lack of legal rules (see *Nature* **404**, 692; 2000). More than 600 geneticists, physicians, ethicists, theologians and women's and disabled persons' rights activists took part.

At a concluding press conference, Fischer called the symposium a great success. "It was extremely important for us all to meet our opponents and listen to them," she said. "Now I can see much clearer the pros and cons of the most recent biomedical opportunities."

Fischer, a member of the Green party, had said earlier that she was keen to start a broad discussion on both medical opportunities and the necessary ethical boundaries of biomedical research. Especially in the light of a possible revision of Germany's ten-year-old embryo protection law, which would turn it into a more comprehensive law on reproductive medicine.

The current law outlaws research on *in vitro* stem cells from human embryos and pre-implantation diagnostics. It has come under fire from many scientists, physicians and legal experts as being outdated, given the speed of advances in the life sciences and the less restrictive legislation in force elsewhere in Europe and the United States.

Presentations to the symposium suggested that although controversy remains over issues such as genetic 'selection' and eugenics, extreme positions are on the wane. But sharp divisions remain over whether advances in biomedicine and genetics are a benefit or a threat to society.

Fischer said that she personally would prefer to maintain tough restrictions on the applications of such techniques, on the basis that pre-implantation diagnostics and genetic screening could lead to a gradual 'devaluation' of sick people and discrimination against the disabled.

Dietmar Mieth, head of the Centre for Ethics in Science at the University of Tübingen and a member of the federal health ministry's ethics commission, says that although



Fischer: encouraging debate over stem-cell research.

scientific freedom is guaranteed in Germany's post-war constitution, this freedom cannot be interpreted as absolute, particularly in the ethically sensitive area of biomedicine.

This view is shared by Jens Reich, a bio-informaticist at the Max Delbrück Centre for Molecular Medicine in Berlin and the

Greens' 1994 candidate for federal president.

Mieth argues that the embryo protection law should be retained. But he would like it supplemented by additional legislation on genetic diagnostics, which would include detailed regulations on pre-implantation diagnostics, and take into account the results and medical implications of the Human Genome Project.

Experts on both sides agreed that if Germany maintains its comparatively restrictive attitude to areas of biomedicine relating to human embryos, the result will be a form of medical and scientific 'tourism' as researchers seek less onerous conditions in neighbouring countries.

But Mieth rejects the argument that excessive restrictions could cut Germany off from frontline biomedical research in areas such as stem-cell research. "Concern and caution are not German quirks," he says. "Stem-cell research is equally controversial in France and in the UK."

The symposium also raised questions about the value of academic and practical ethics in the decision-making process. Kurt Bayertz, a philosopher at the University of Münster, argued that consensus can only be achieved on a political level. Ethics have no affinity to consensus, he said, in the context of problems such as the moral status of the human embryo. ■

Survey nets volcanic prize

Peter Pockley, Sydney

The largest volcanic chimney ever to be recovered from the sea floor is being brought ashore by an international team working for the Commonwealth Scientific and Industrial Research Organisation. Caught by the marine research vessel

Franklin during an expedition focusing on mineral formation in the Western Pacific Ocean, the chimney (inset) is 2.7 m high, 70–80 cm across and weighs one tonne. It was snagged by a dredging net during a routine survey of the Eastern Manus Basin in the Bismarck Sea.

The research team also witnessed remarkable island-building activity as the marine volcano Kavachi erupted for the first time in nine years (below). Kavachi, near the Solomon Islands, produced eruptions 800 metres high every five minutes, ejecting ash and incandescent blocks of lava up to 70 metres above sea level.



French scientists turn to journals in English

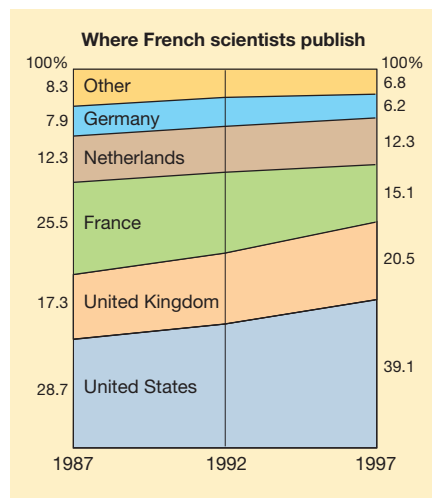
Declan Butler, Paris

French researchers appear to be making the most of a 1994 court ruling allowing them to publish in languages other than French, according to a study by the Paris-based Observatoire des Sciences et des Techniques (OST). The study shows that French scientists are increasingly publishing in English-language journals.

The number of articles by French authors published in English-language journals has leapt between 1987 and 1997. The proportion of these published in the United States increased to more than a third, from 28.7 per cent to 39.1 per cent (see graph).

A similar pattern occurred in the proportion of articles appearing in journals published in Britain. In contrast, the number of French-authored articles appearing in journals published by Dutch publishing houses has remained flat, while French scientists have reduced their publications in German, Swiss and Japanese journals.

French scientists have always been ill at ease with the role of English as the international language of science. In 1994, a court intervention prevented the enactment of a law requiring nearly all publicly funded research to be published in French. The court



ruled that the law was contrary to the “freedom of communication and expression in teaching and research”.

Reflecting this increased willingness (and desire) to publish abroad, the overall proportion of French-authored papers in French journals dropped by 41 per cent over the decade. This also reflects an increase in the number of scientists from overseas who publish in French journals. Japanese

scientists, for example, doubled the number of their papers in French journals over the period, while British-authored papers have increased by around a third.

The data collected by the OST confirm a growing domination of the journals market by the United States, the United Kingdom and the Netherlands.

The European Union as a whole accounts for 1,810 of the 3,681 journals indexed by the Philadelphia-based Institute for Scientific Information, ahead of the United States, which produces 1,442 of the journals in the database. But within Europe, France is a minor publisher, with just 96 journals.

This compares with 776 published in the United Kingdom, 399 in the Netherlands — a reflection of the role of large publishing houses such as Elsevier — and 288 in Germany (Japan produces 122). France's share of the world journal output fell from 3.1 per cent in 1987 to 2.6 per cent in 1997.

The OST points out that journals are just one of the many means of scientific communication, along with conferences, e-prints and electronic journals. Future studies, it emphasizes, will need to look beyond traditional journals to gain an accurate impression of publishing trends. ■

World's academies seek a sustainable future

Robert Triendi, Tokyo

Governments were urged last week to increase their investment in areas of basic science and technology related to sustainable development. This was the message to emerge from a four-year study of sustainability by the world's scientific academies, whose results were presented at a meeting in Tokyo.

It was also agreed that the secretariat for the InterAcademy Panel on International Issues (IAP), which set up the study and has up to now been run from the Royal Society in London, will move to the Third World Academy of Sciences (TWAS) in Trieste, Italy. A spokesman for TWAS said the academy was "delighted".

The Italian government, which already provides backing for both TWAS and the International Centre for Theoretical Physics, is expected to support the housing of the IAP secretariat in a villa being made available by regional authorities.

A statement at the end of the meeting on the need for more research into sustainable development has been criticized by some for being too vague, but others say the wording was deliberately kept open. "The details will have to be filled out over the next year," says



Bonjour Trieste: the Italian port will become the new seat of the InterAcademy Panel.

Sherwood Rowland, foreign secretary of the US National Academy of Sciences and outgoing co-chairman of the IAP.

According to Yves Quéré, foreign secretary of the French Academy of Sciences and one of the two new co-chairmen (the other is Eduardo Krieger, president of the Brazilian Academy of Sciences), one goal of the IAP will be to organize projects around 'clusters' of academies sharing common interests. He cites an existing collaboration between French and Swedish scientists on mother/infant health. Other topics could include school science education and the

biodiversity of tropical rainforests.

The IAP will work closely with the new InterAcademy Council, set up in Tokyo. Based at the Royal Netherlands Academy of Arts and Sciences in Amsterdam, the council will provide scientific expertise to international organizations such as the United Nations and the World Bank (see *Nature* 405, 268; 2000).

In contrast, the IAP will continue primarily to help its members to fulfil their consultative and administrative functions within their home countries.

► <http://www.nationalacademies.org/iap>

DENNIS MARISCO/CORBIS

US National Ignition Facility criticized by fellow laboratory

Washington Deep divisions between US nuclear weapons laboratories over the future of the troubled National Ignition Facility (NIF) fusion project spilled into public view last week. A senior official at the Sandia laboratory in New Mexico issued a statement questioning "what is a reasonable additional investment in the NIF".

The statement, from Sandia vice-president Tom Hunter, was quickly withdrawn by the laboratory and condemned by Bill Richardson, the energy secretary, who pledged to ignore it. But its contents confirmed growing concerns at Sandia and the Los Alamos National Laboratory, also in New Mexico, that cost overruns at the NIF could endanger other parts of the nuclear weapons research programme.

An unpublished report by the General Accounting Office says that the NIF's total construction cost, which was supposed to be \$1.2 billion, could reach as much as \$3.6 billion. Senator Pete Domenici (Republican, New Mexico) is strongly resisting efforts to help pay for the cost overruns from the budget of Sandia or Los Alamos.

Phenomenal result for Durham in particle physics

London Britain's University of Durham has beaten ten rivals from across the country to house a new £12 million (US\$17.6 million) Institute for Particle Physics Phenomenology. The institute will have a new building and 50 staff and research students, double the existing number. Durham is already home to one of the United Kingdom's largest groups of theoretical particle physicists.

Phenomenology is the bridge between theory and experiment in particle physics; the institute will test theories of the fundamental forces of physics and the structure of matter. The institute will be supported, for at least ten years, by the Particle Physics and Astronomy Research Council and the university. Its first director is James Stirling.

Human gene map on schedule for autumn

Hay-on-Wye, Wales One of the leading figures in the Human Genome Project said last week-end that the publication of the full draft of the genome sequence was now expected "some-time this [autumn]". Eric Lander, director of the Whitehead Institute at the Massachusetts Institute of Technology, also said that the public was being "misled" by media descriptions of a race to the finishing line, stressing that much of the data produced by the



Under fire: Edward Moses, head of the troubled National Ignition Facility (see left).

sequencing project are "already being used in laboratories throughout the world".

Delivering a *Nature*-sponsored lecture at the annual Hay-on-Wye literary festival, Lander, whose laboratory is one of the main US sequencing centres, also called for an international moratorium on the germline modification of human beings. Although he noted that this should not necessarily be a permanent ban, he said that he "would hate to wake up one morning to find that someone had produced a genetically modified child".

Ethics invoked to revise euthanasia drug patent

Munich The European Patent Office (EPO) came under fire last week over a patent on a euthanasia drug for mammals, granted to Michigan State University, that included possible application on humans. Several German, European and US pressure groups opposing human euthanasia challenged the patent on ethical grounds, fearing that it could encourage human euthanasia.

The EPO's opposition board accepted the ethical objections, and has restricted the patent's validity to lower mammals. According to an EPO spokesman, this is the first time that a patent has been modified solely because of ethical concerns.

Budget setback for US science agencies

Washington Hopes for a major increase in funding for the US National Science Foundation (NSF) suffered a setback last week, as appropriations subcommittees in the House of Representatives passed their initial appropriations bills for key research agencies. Science advocates remain optimistic that each of these agencies will do better than the initial House offering when the budget process for the 2001 fiscal year is completed in the autumn.

The subcommittee with jurisdiction over the NSF offered it an increase of \$167 million, or 4 per cent, against the 17 per cent proposed by President Bill Clinton, and

granted NASA an increase of \$112 million, or less than 1 per cent, against 4 per cent proposed by Clinton. And although the \$1 billion increase passed by another House subcommittee for the National Institutes of Health surpassed Clinton's proposal, it fell short of the \$2.7 billion sought by biomedical research advocates and granted earlier by the Senate appropriations subcommittee.

TV programme pre-empts German research on bees

Munich German researchers at the University of Jena have been embarrassed by a premature television report describing their as yet unpublished research. According to the report, the work provides the first evidence of the horizontal transfer of a herbicide resistance transgene from plants to the gut bacteria of bees.

The day after the report was broadcast — without the scientists' approval — the university issued a press release about the four-year study, which monitored bees feeding on pollen from genetically engineering maize and rape. The researchers were quoted as emphasizing that the transgene was only detected in a small number of bees, and that these were not harmed by the transfer.

New commission to focus on US oceans

Washington The Pew Charitable Trusts has set up a high-powered commission led by Todd Whitman, the governor of New Jersey, and Leon Panetta, former White House chief-of-staff, to assess America's oceans and recommend actions to assist marine conservation.

Scientists on the Pew Oceans Commission will include Jane Lubchenco of Oregon State University and Charles Kennel, director of the Scripps Institution in California. The commission promises to conduct "an aggressive schedule" of public hearings and to produce interim reports within 18 months. It will draw on Pew's considerable resources to boost the public profile of marine conservation issues in the United States.

Correction

As a result of cuts made during editing, the News Feature on gene silencing (*Nature* 404, 804; 2000) failed to mention that post-transcriptional gene silencing in plants was co-discovered by Joseph Mol and his colleagues at the Free University of Amsterdam. Researchers led by Jan Kooter in Mol's lab were also the first to demonstrate the post-transcriptional nature of the phenomenon. Interested readers can find a full review of the history of gene silencing in plants in *Trends in Plant Science* (4, 340–347; 1999).

Requiem for an observatory

One of NASA's most productive satellites is about to meet a fiery end. Henry Bortman assesses the achievements of the Compton Gamma-Ray Observatory, and considers the missions that will follow in its footsteps.

Early in the morning of 4 June, a swath of the central Pacific Ocean some 4,000 kilometres southeast of Hawaii will be kept clear of shipping and aircraft — and for good reason. If all goes to plan, 17 tonnes of satellite will come hurtling out of orbit, disintegrate in the atmosphere, and rain down as a series of fireballs over a belt roughly 1,500 kilometres long by 25 kilometres wide.

It will be a sad day for science. The Compton Gamma-Ray Observatory (CGRO), which is being scuttled by NASA to avoid the remote possibility that it might make an uncontrolled descent over an inhabited area, has been a stellar performer. In the nine years since it was put in orbit by the Space Shuttle Atlantis, the CGRO has transformed astronomers' understanding of some of the most intriguing objects in the Universe.

And although other, more powerful gamma-ray instruments will be launched over the next few years, Compton's loss will leave gamma-ray astronomers helpless should an important high-energy astrophysical event occur during the hiatus. "For a few years we'll be just blind," says Isabelle Grenier of the University of Paris VII.

Great expectations

The CGRO's demise also marks a turning point in NASA's philosophy. Compton is one of the four 'Great Observatories' — astronomical satellites of unprecedented size and performance designed to study the sky across a range of wavelengths from infrared to gamma-rays. The group also includes the Hubble Space Telescope, which recently celebrated its tenth anniversary in space; the Chandra X-Ray Observatory, launched last July; and the Space Infrared Telescope Facility (SIRTF), scheduled for launch in December 2001.

The plan was to overlap the operation phases of each satellite to let astronomers use

We need more specialized missions that focus on a particular problem and go after it with a vengeance.



Vision on: launched nine years ago, the Compton Observatory (main picture) has allowed Earth-bound instruments such as ROTSE (inset) to study stellar gamma-ray bursts with new accuracy.

different wavelengths simultaneously to observe objects, allowing them to gain insights that would be impossible to achieve using just a narrow wavelength band. But each observatory has suffered a series of delays — the programme's schedule was thrown into disarray from the start after NASA's shuttles were grounded following the 1986 Challenger disaster. And in losing the CGRO after Chandra has been in orbit for less than a year, and before SIRTF has even left the ground, the goal of overlapping observations at all wavelengths has clearly been missed.

Now they have got their hands on the data that the CGRO has amassed, however, most astronomers aren't too concerned. They are still committed to the idea of studying objects at various wavelengths simultaneously. But, with the exception of near-infrared astronomy, where plans for the \$1 billion Next Generation Space Telescope to succeed Hubble are well advanced, the consensus is that this will in future be best done using smaller satellites focused on particular scientific goals.

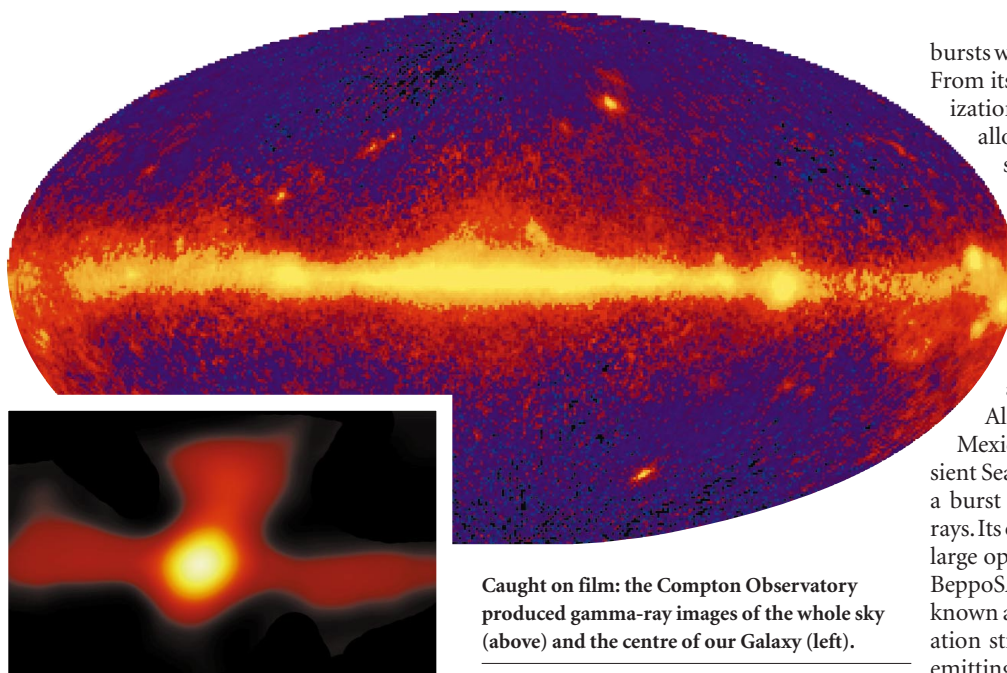
"We're not doing Great Observatories any more, because it puts too many eggs in one basket, frankly," says Alan Bunner, director of NASA's Structure and Evolution of the Universe programme. "What's needed next

in this field of astrophysics are more specialized missions that focus on a particular capability or a particular problem and go after that with a vengeance."

Bursting on the scene

Although astronomers are excited by the possibilities offered by the missions on the drawing-board, they view the CGRO's demise with regret. The observatory may not have achieved the fame of Hubble, but that has little to do with its scientific significance. Compton isn't a household name simply because its instruments couldn't provide stunning pictures that could be splashed across the world's newspapers. But with the observatory facing its doom, astronomers are keen to sing its praises.

"Probably the most important discovery had to do with gamma-ray bursts," says Neil Gehrels, project scientist for the CGRO, based at NASA's Goddard Space Flight Center in Greenbelt, Maryland. These flashes of radiation last from a few milliseconds to a few minutes, and were first recorded in the 1960s by spy satellites searching for evidence of nuclear weapons tests. Prior to the Compton observatory, there was not much information on the bursts. "But what little there was had convinced the community completely — I



Caught on film: the Compton Observatory produced gamma-ray images of the whole sky (above) and the centre of our Galaxy (left).

mean almost to a person — that since they're such bright flashes of gamma-rays, they must be something occurring in our own region of the Milky Way," says Gehrels.

Within months of its launch, however, Compton was overturning that view. One of its main instruments is the Burst and Transient Source Experiment (BATSE). This allowed astronomers to monitor the distribution of the gamma-ray bursts across the sky, and soon showed that they come at a rate of about one per day from all directions¹. The fact that the bursts aren't concentrated in the plane of our Milky Way Galaxy suggested that they come from great distances away — which in turn indicated that their sources must be the most energetic explosions in the Universe. "This changed the subject of gamma-ray bursts from a little exotic side topic to a major discipline in astronomy," says Gehrels.

The CGRO also pioneered studies of a class of objects dubbed 'blazars'. These are quasars, or distant active galaxies, from which huge jets of matter stream out at almost the speed of light. "Even before the CGRO, looking at radio data and some optical data, there was this odd set of quasars that seemed to have different properties," says Gehrels. "Instead of being steady objects, the light coming out of them would vary from day to day fairly rapidly."

Web links

CGRO ♦ <http://cossac.gsfc.nasa.gov>
 Great Observatories ♦ <http://sirtf.jpl.nasa.gov/sirtf/Mission/Family/greatobs.html>
 HETE-2 ♦ http://international.gsfc.nasa.gov/International/Missions/Hete_2/Hete_2.html
 Swift ♦ <http://swift.sonoma.edu>
 GLAST ♦ <http://www-glast.stanford.edu>
 INTEGRAL ♦ <http://astro.estec.esa.nl/SA-general/Projects/Integral/Integral.html>

The nature of these quasars was clarified by an instrument aboard Compton called the Energetic Gamma-Ray Experiment Telescope (EGRET), which revealed that they were also emitting gamma-rays²⁻⁴. Because astrophysicists know that high-energy gamma-rays are generated when particles are accelerated to very high speeds, explains Gehrels, "we were able to show in a solid way that these were really coming from jets flowing out".

Puzzling evidence

Some of Compton's discoveries, however, have provided more questions than they have answers. Using EGRET, astronomers have documented around 170 point sources of gamma-rays that defy explanation^{5,6}. These appear in the plane of our Milky Way, and so are thought to lie in the main disk of our Galaxy, a kiloparsec or more away.

Earlier this year, a team led by Gehrels reported the existence of a smaller group of fainter sources lying in a nearby star-forming zone called the Gould Belt, 100–400 parsecs from our Solar System⁷. The nature of both classes of sources remains a mystery. "We don't have a clue what they are," admits Bunner. "They don't show up — at least no one has been able to find the counterparts — at other wavelengths."

Coordinating Compton's observations with those at other wavelengths has proved invaluable in identifying other gamma-ray sources, including the transient bursts. Within seconds of detecting a burst, the CGRO can transmit its coordinates — admittedly not very well localized — to other space- and ground-based observatories.

This approach really came into its own when Compton was teamed up with the Italian–Dutch satellite BeppoSAX. This craft observes at gamma-ray and X-ray wavelengths, and can pinpoint the origin of the

bursts with greater precision than Compton. From its launch in 1996, BeppoSAX's localization of bursts recorded by Compton allowed astronomers to train their telescopes on the area of sky from which the bursts came, and so watch their afterglow^{8,9}. Measurements of the redshift of the optical counterparts of the bursts confirmed that they were coming from distant galaxies⁹.

Most significantly, on 23 January 1999, an instrument at the Los Alamos National Laboratory in New Mexico, called the Robotic Optical Transient Search Experiment (ROTSE), observed a burst while it was still emitting gamma-rays. Its observations¹⁰, together with those of large optical telescopes guided by data from BeppoSAX¹¹⁻¹⁵, suggested that the source, known as GRB990123, was beaming its radiation straight in our direction, rather than emitting it in all directions.

The hole story

If this pattern is typical of gamma-ray bursts, it means that the total energy of the explosions responsible need not be as high as astrophysicists thought. This fits with the theory that at least some of the bursts are caused when a supermassive star goes supernova. According to this 'collapsar' model, material blown away by the explosion will collapse back onto the black hole left behind, and then be emitted as highly energetic jets from its poles. "The follow-up at all wavelengths has been the key to cracking this problem," says Stan Woosley of the University of California at Santa Cruz, one of the proponents of the collapsar model.

The same argument applies to other gamma-ray sources. "Once it's been possible to look at objects across the whole spectrum, it's become pretty blindingly obvious that that was the way to study them," says Ken Pounds, a high-energy astrophysicist at the University of Leicester, and formerly chief executive of Britain's Particle Physics and Astronomy Research Council. "If you take something like the nucleus of an active galaxy, then you will find that it puts out pretty much equal amounts of power in all



Sunblind: the Compton Observatory's early demise will hamper studies of solar flares.

news feature

parts of the spectrum: radio, infrared, optical, ultraviolet, X-rays and even gamma-rays. To understand these objects, you really do have to look at them in all these different ways. Otherwise, it would be like trying to draw a picture of a human being by only looking at an arm or a leg."

Given this, you might expect that astronomers working on the Great Observatories programme would be distraught at losing the CGRO before Chandra has really hit its stride, and with SIRTf still waiting in the wings. But most are relatively sanguine. "It would have been nice to have Chandra and the CGRO up together for longer," says Gehrels. "Chandra is following up, making observations of gamma-ray bursts, studying the X-ray afterglow in a way that we've never been able to before. But if you look at the overall science programme of Chandra, it's a fairly small fraction."

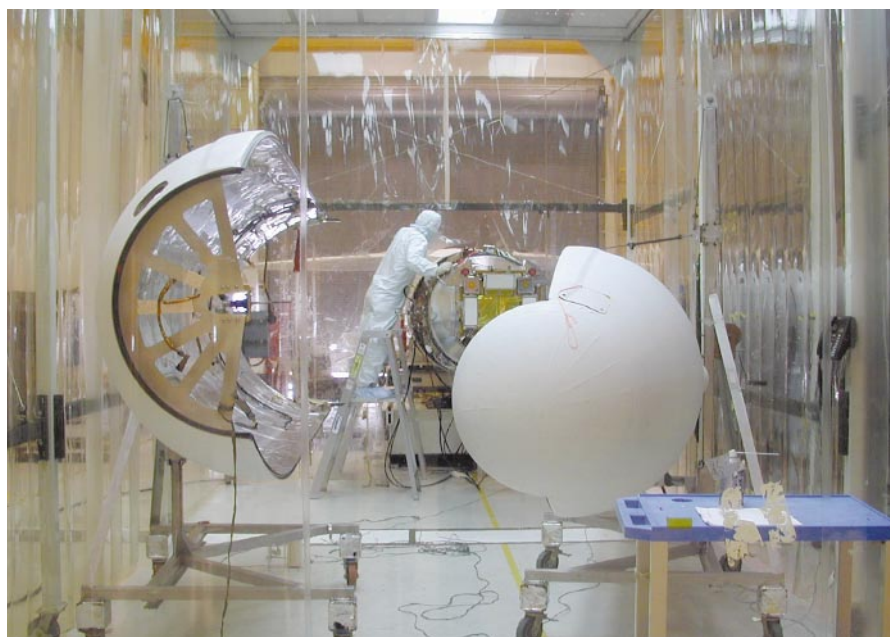
Sharpshooting satellites

Indeed, most astronomers are now looking forward to a series of smaller, more specialized gamma-ray missions that should fill in the details of the broad-brush picture painted by Compton. The High Energy Transient Explorer-2 (HETE-2), for instance, is a gamma-ray burst detector set for launch by a NASA-led international team in July. Although less sensitive than Compton's BATSE instrument, HETE-2 will be far better at pinpointing the location of the bursts it does detect, making it easier to coordinate observations at other wavelengths.

"I'm absolutely sure that, when HETE-2 goes up, you'll see a huge revitalization of the gamma-ray burst field," says Shri Kulkarni, an astrophysicist at the California Institute of Technology in Pasadena. HETE-2 should be followed in 2003 by Swift, a NASA mission that will combine a good ability to pinpoint the bursts with instruments that are five times as sensitive as Compton's.

The Gamma-ray Large Area Space Telescope (GLAST) is scheduled for launch by NASA in 2005. It will focus on studying the highest-energy end of the gamma-ray spectrum. With 30 times Compton's sensitivity for gamma-rays with energies in the giga-electron-volt range, and 10 times the angular resolution, GLAST should help to unlock the secrets of gamma-ray objects whose sources have not yet been identified. The European Space Agency is also getting in on the act with its International Gamma-Ray Astrophysics Laboratory (INTEGRAL). This mission, which should reach orbit in April 2002, will provide spectra for gamma-ray sources within our own Galaxy.

Despite the general feeling that these focused missions are the way of the future, some researchers are upset about NASA's decision to sacrifice Compton. Jim Ryan of the University of New Hampshire in Durham is co-principal investigator for an



Ray of hope: when it's launched this July, HETE-2 (the High Energy Transient Explorer-2), here undergoing testing, will be able to pinpoint the locations of elusive gamma-ray bursts.

instrument called COMPTEL, or the Imaging Compton Telescope. He has been circulating a letter among astrophysicists urging the US Congress to suspend NASA's de-orbiting plan and to start an independent investigation of the safety risks.

In-between days

NASA decided to bring the craft down after a gyroscope failure last December put it one more gyro malfunction away from a possible uncontrolled re-entry. Given that some sections of the huge satellite are bound to reach the Earth's surface, this carries a one in a thousand risk of causing human casualties. However, alternative control techniques that don't rely on gyroscopes might reduce that chance to one in 4 million — a risk some astronomers feel is worth taking. "I bet you a nickel to a dime that there's more risk to the public in the launch of a shuttle than there is in bringing down the CGRO," says Ryan.

Ryan is particularly frustrated about lost opportunities to observe gamma-ray emissions from the Sun. "We're in the maximum of solar activity right now, and we've just begun to start acquiring some terrific solar flare data for which we've been waiting 10 years," he says. NASA's High Energy Solar Spectroscopic Imager (HESSI) mission, designed to observe solar flares at X-ray and gamma-ray wavelengths, was supposed to have been launched in July, and could have taken over solar observations from Compton. But it was damaged in March during bungled vibration tests, and is now unlikely to fly before January 2001.

Bunner is sympathetic to Ryan's concerns. "People hate the thought of turning off

any observatory, especially one that's still functioning," he says. But, according to the safety framework within which NASA has to operate, Bunner says that the agency had little choice but to bring Compton down. NASA officials did consider sending a shuttle mission to repair the ailing spacecraft, but that was deemed too expensive — and too dangerous for the astronauts who would have to perform the spacewalks.

Most astronomers seem willing to accept this logic, although they're aware that Compton's demise means that the lights are about to go out for a while on the gamma-ray sky. "I think that we'll get by with a two- or three-year gap, provided there isn't an extremely rare event that needs a gamma-ray mission to follow it properly," says Pounds. "For instance, if there were to be a supernova explosion in our Galaxy — a once-in-a-hundred-year event — and we missed seeing it in part of the spectrum where it was probably going to be very interesting, then that would be a considerable loss." From 4 June, high-energy astrophysicists will just have to keep their fingers crossed. ■

Henry Bortman is a freelance writer in San Francisco.

1. Meegan, C. A. *et al. Nature* **355**, 143–145 (1992).
2. Hartman, R. C. *et al. Astrophys. J.* **385**, L1–L4 (1992).
3. Mattox, J. R. *et al. Astrophys. J.* **476**, 692–697 (1997).
4. Mukherjee, R. *et al. Astrophys. J.* **490**, 116–135 (1997).
5. Mukherjee, R., Grenier, I. A. & Thompson D. J. in *Proc. 4th Compton Symp.* (eds Dermer, C. D., Strickman, M. S. & Kurfess, J. D.) 394–406 (American Institute of Physics, New York, 1997).
6. Hartman, R. C. *et al. Astrophys. J. Suppl.* **123**, 79–202 (1999).
7. Gehrels, N., Macomb, D. J., Bertsch, D. L., Thompson, D. J. & Hartman, R. C. *Nature* **404**, 363–365 (2000).
8. van Paradijs, J. *et al. Nature* **386**, 686–689 (1997).
9. Metzger, M. *et al. Nature* **387**, 878–880 (1997).
10. Akerlof, C. *et al. Nature* **398**, 400–402 (1999).
11. Kulkarni, S. *et al. Nature* **398**, 389–394 (1999).
12. Galarmat, T. *et al. Nature* **398**, 394–399 (1999).
13. Castro-Tirado, A. J. *et al. Science* **283**, 2069–2072 (1999).
14. Hjorth, J. *et al. Science* **283**, 2073–2075 (1999).
15. Andersen, M. I. *et al. Science* **283**, 2075–2077 (1999).

To conserve rainforest, we have to help local people live sustainably

Sir— We disagree with Gullison *et al.* who argued in their Commentary “Marketing” species conservation¹ that the Convention on International Trade in Endangered Species (CITES) should be restructured to protect mahogany (*Swietenia macrophylla*). In our view, CITES has been effective in regulating international trade of some species endangered with biological extinction.

Although logging can have catastrophic impacts on forest ecosystems, few internationally traded tropical hardwood species are threatened with extinction. Of the ten most important tropical hardwoods imported to the United Kingdom, only *S. macrophylla* is listed by CITES (in its Appendix III). Logging has been accused of causing the local extinction of mahogany². However, we have recently conducted 100 per cent inventories of mahogany trees in three, conventionally logged areas in two different forests in Brazil (unpublished data). Mahogany survived logging in all three areas (Table 1), but few trees are yet of commercial size. In these forests, mahogany is not threatened with biological extinction, but further exploitation is unlikely to be profitable for many years.

Although forests are logged tens to hundreds of kilometres from centres of the timber trade³, large areas of mahogany's vast natural range⁴ remain remote and unexploited by commercial logging. Mahogany is widely planted around the tropics, and there are now promising solutions to its silvicultural problems^{5,6}. Mechanisms other than CITES need to be found to curb destructive logging in tropical rainforests and to ensure the sustainable management of commercial species, such as mahogany, which are not yet endangered.

Gullison *et al.* argue that the rate at which sustainable forest-management practices are being adopted is too slow to save many species. They recommend that organizations in the developed countries should buy and protect tracts of tropical forest as a safety net if other measures continue to be ineffective in areas of rapid forest decline. Although this strategy may be effective in the short term, we do not believe that it is viable in the long term. Unless alternatives can be found to the unsustainable exploitation and clearance of forest, such protected areas will become increasingly isolated and vulnerable to illegal logging, fire and occupation.

Illegal logging is widespread in the tropics⁷. Even legally logged forests are often repeatedly and illegally logged for residual timber, unless they are well protected. Illegal occupation of forest is also common. Many

Table 1 Density of mahogany in two forests in Pará state, Brazil

Location	Years since logging	Trees per ha	Stumps per ha
Marabá	14	0.06	0.03
Rio Maria	10	0.66	1.29
Rio Maria	7	0.40	0.45

Density of *Swietenia macrophylla* stumps and trees ≥ 15 cm diameter in three 300-ha logging coupes in two different forests

tropical rainforests lie adjacent to burgeoning human populations. In Pará, Brazil, landless people have recently occupied private forest reserves, government-owned research forests and privately owned production forests. Who will pay for long-term protection as the agricultural frontier advances? Some Pará landowners have hired gunmen to protect their property, a strategy that is unlikely to be palatable to many conservation organizations.

Although it may be cheap to lease logged forests, Gullison *et al.* ignore the high cost of long-term protection. Leasing is not an option in the Old World tropics where forests are state property. The leasing proposal would not slow the rate at which new forests are penetrated by logging roads, ‘mined’ for mahogany and thereby made accessible to land conversion.

Alliances between conservationists and local communities have rarely proved successful in the long term. For example, the Projeto de Estudio para el Manejo de las Areas Silvestres de Kuna Yala in Panama initially received funding of US\$1.2 million and had a high international profile. However, within a few years the money had run out as there was no planning for the long-term financial sustainability of the project and no systematic fund-raising⁸.

We believe that the only viable strategy for conserving large areas of tropical rainforest is to find sustainable livelihoods for rural populations and to control destructive logging through marketplace initiatives. Sustainable management of tropical rainforests, including the control of illegal logging, fire and land occupation, is therefore a necessary precondition for successful conservation. Totally protected areas are doomed if they ignore the surrounding land-use issues.

Sustainable exploitation is increasingly being forced by consumer pressure. Europe will now only import tropical timbers certified as sustainably produced by bodies such as the Forest Stewardship Council, an international non-profit organization.

Rather than buying reserves, conservation bodies may be better advised to spend their money in offsetting the high costs of such certification and in educating consumers to pay premium prices for credibly certified timber.

S. Jennings*, N. D. Brown*, T. C. Whitmore†, J. N. M. Silva‡, J. do C. A. Lopes‡, A. M. V. Baima‡

*Oxford Forestry Institute, University of Oxford, Oxford OX1 3RB, UK

†Department of Geography, University of Cambridge, Cambridge CB2 3EN, UK

‡Embrapa Amazônia Oriental, CEP:66.005 100, Belém, Pará, Brazil

1. Gullison, R. E., Rice, R. E. & Blundell, A. G. *Nature* **404**, 923–924 (2000).
2. Snook, L. K. & MacLellan, A. *Bot. J. Linn. Soc.* **122**, 35–46 (1996).
3. Verissimo, A., Barreto, P., Tarifa, R. & Uhl, C. *Forest Ecol. Mgmt* **72**, 39–60 (1995).
4. Styles, B. T. in *Flora Neotropica Monograph* (eds Pennington, T. D., Styles, B. T. & Taylor, D. A. H.) 389–406 (New York Botanic Garden, New York, 1981).
5. Mayhew, J. E. & Newton, A. C. *The Silviculture of Mahogany* (CAB International, Wallingford, 1998).
6. Neill, D. A. & Ravelo, N. in *The Genus Inga Utilization* (eds Pennington, T. D. & Fernandes, E. C. M.) 141–150 (Royal Botanic Gardens Kew, London, 1998).
7. Rodan, B. D., Newton, A. C. & Verissimo, A. *Environ. Conserv.* **19**, 331–342 (1992).
8. Chapin, M. in *From Principles to Practice: Indigenous Peoples and Biodiversity Conservation in Latin America* (eds Gray, A., Parellada, A. & Newing, H.) 240–278 (IWIGA, Copenhagen, 1998).

Impact factors aren't relevant to taxonomy

Sir— Evaluating research by means of ISI impact factors — those determined by the Institute for Scientific Information in Philadelphia — gives rise to obvious problems in the case of basic biodiversity research. Valdecasas *et al.* in Correspondence (*Nature* **403**, 698; 2000) suggested that one cause is the low citation rate of taxonomic articles compared to other fields.

A more important reason, however, is the inapplicability of this index to basic biodiversity research on principle.

The impact factor generally underestimates the number of citations because of the limited number of scanned journals. In fields such as molecular genetics, neuroscience or cancer research — where, as Bradford's law states, most of the relevant work is published in a few core journals — this selection does not severely affect the comparability of impact factors.

However, Bradford's law does not apply to taxonomy and other areas of basic biodiversity research. Generally, taxonomic papers contain details of nomenclature, which must be considered and discussed whether the paper is excellent or poor. Similarly, reports of flora and fauna records need to be debated in further studies before being accepted or rejected. Quality and relevance are independent parameters.

Therefore, it is impossible to classify taxonomic or ecology journals as more or less important. They can only be classified as of high or low quality, which does not affect the number of citations. Many natural history journals contain taxonomic or similar articles. The library of London's Natural History Museum alone holds 11,000 serial titles,

most of which come under this category, and each may contain relevant information. About 1,000 entomological journals are held by the Natural History Museum's entomology library, but only 65 are covered by the expanded Science Citation Index (SCI). Because some 93 per cent of potentially relevant journals are not considered, it is good luck rather than a sign of good quality to be included in the SCI.

The impact factor is a powerful and welcome tool in some disciplines. In basic biodiversity research, however, it is not applicable. Qualified referees must evaluate the scientific work itself.

Frank-Thorsten Krell

Entomology Department, Natural History Museum, Cromwell Road, London SW3 5BD, UK

We have moral authority to apologize for the past

Sir — Your editorial "Hollow apologies should be avoided" (*Nature* **403**, 813; 2000) noted that Hubert Markl, president of the Max Planck Society (MPS), "is right to resist pressure to apologize" for the participation of German scientists in Nazi experiments.

Subsequently, Bernd Wirsing of the MPS wrote (*Nature* **404**, 222; 2000) that "President Markl has already publicly apologized", although reference to the MPS website reveals that this apology is weakened somewhat by circumlocutory equivocation. The issue of apology by proxy remains relevant in this context and in others.

The Pope and President Wahid of Indonesia have recently apologized for the sins and crimes of their predecessors, but the Australian prime minister, John Howard, has consistently refused to apologize to Aboriginal Australians for generations of abuse and neglect by governments. Howard's rationale is much the same as that put forward in *Nature*'s editorial: the successors of those who committed crimes in the past do not have the moral authority to apologize on their behalf.

Such caution is mistaken. Consider an analogy: if one were enfeebled or incapacitated, it would be reasonable to place one's affairs in the hands of an attorney who would then assume one's legal powers and exercise them in one's stead for one's good.

Markl is the president of the organization that fully took over the members of its predecessor, the Kaiser Wilhelm Society, under whose auspices experiments on human beings took place. Markl is in the position of acting for a morally incapable predecessor. The Kaiser Wilhelm Society was morally deficient in using children and other Nazi victims in its experiments. Recalcitrant

members might well refuse to acknowledge the wrongfulness of their actions or to ask for forgiveness. So indeed, might many of our forebears and contemporaries in this country.

But the MPS and Australia's leaders are not deficient in that respect. They can act as the rightful exercisers of moral authority here, and they ought to do so precisely on the grounds that Markl advances: that those who performed experiments on humans or so egregiously treated Aborigines were deficient in their moral understanding and certainly incapacitated from making a genuine apology.

Finally, it is wrong to argue that an apology needs to be made only to the survivors of atrocious policies and practices. It surely needs to be made to all the descendants of those who were affected — survivors or the murdered — and to the rest of humanity.

Damian Grace*, John Carmody†

*Faculty of Arts and Social Sciences, †Faculty of Medicine, University of New South Wales, Sydney, New South Wales 2052, Australia

Offside researchers score an own goal

Sir — In an astounding analysis of the failure of assistant referees to judge offside situations correctly (R. D. R. Oudejans *et al.* *Nature* **404**, 33; 2000), a group of soccer scientists from — almost inevitably — the Netherlands define the object of their study thus: "In football (soccer), a player is 'offside' if he or she is closer to the goal than the last defender (excluding the goalkeeper) when the ball is passed to them."

However, 'offside' in football depends on an attacker's distance from the goal line (the line forming the boundary of the pitch on that side, on which the defending team's goal is placed) rather than from the goal itself. Moreover, with respect to the offside rule the goalkeeper has the same status as any other player. Could it be that, in reality, assistant referees err so often because they don't know the rules of the game any better than those who conduct scientific studies on the rules?

For the correct rule see <http://www.fifa2.com/scripts/runisa.dll?s7.6750696:gp:919357:67173+refs/laws/law11> — or ask somebody in the nearest pub.

Winfried S. Peters*, Bernd Herkner†, Alex Webb‡, Julian Hibberd‡

*Kinematic Cell Research Group, Goethe University Biocentre, Marie-Curie-Strasse 9, D-60439 Frankfurt am Main, Germany

†State Museum of Natural History, Erbprinzenstrasse 13, D-76133 Karlsruhe, Germany

‡Department of Plant Sciences, Downing Street, Cambridge CB2 3EA, UK

Let Armenia show why it's the place for Sesame

Sir — Organizers of the proposed new research centre for the Middle East, Caucasus, Mediterranean and Gulf region, which will be based around the Sesame synchrotron, are saying it will be built in Jordan (*Nature* **404**, 798; 2000). But the decision has not yet been officially made — and at the Interim Council meeting in Geneva, on 10 and 11 April, both Jordan and Armenia were chosen as likely sites.

I believe it is a violation of fair and objective selection practices for organizers of such an important international site to show this bias towards Jordan before the final decision has been made.

As chairman of the Armenian Committee for Sesame, established by the Republic of Armenia, I suggest that an independent panel of accelerator experts from around the world should visit both sites. They could then make an evaluation on the basis of each country's technical expertise and government commitment, and assess which offers the quickest and most economical way to make the project operational.

Armenia and Jordan have submitted their list of advantages. Let the independent panel review each item carefully and make their recommendation to the Sesame Interim Council and to Unesco. The birth of an international centre such as Sesame deserves this careful and fair process. For more information, please visit our website at www.natureswonder.com/sesame.

J. S. Hovnanian

J. S. Hovnanian and Sons, Mount Laurel, New Jersey 08054, USA

Giving Arabs credit for Persian works (again)

Sir — I'm afraid the Book Review of Jean-Luc Chabert's *A History of Algorithms* (*Nature* **403**, 703; 2000) contains the same frequently made mistake pointed out by your correspondent S. Khochbin (*Nature* **405**, 14; 2000) in another recent article.

The reviewer includes the Persian mathematician and scholar Nasir al-Din al-Tusi among Arab mathematicians. Nasir al-Din al-Tusi did write in Arabic (as well as Persian), but this is like saying Sir Isaac Newton was not English, because he wrote in Latin!

Nasir al-Din al-Tusi was born in Tus, a city in the northeast part of Iran, in the province of Khorasan, quite far from any Arab populations at the time, or even now.

Payman Arabshahi

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, MS 238-343, Pasadena, California 91109, USA

Time for voices to be raised

Scientists must become more involved in controversial public debates.

Derek Burke

Attitudes to genetically modified (GM) foods and organisms so far have tended to be black and white — GM is either all good or all bad. Despite guidance from scientists (see <http://www.dti.gov.uk/ost/ostbusiness/index.htm>), media coverage in the United Kingdom is best captured by the phrase 'Frankenstein food'.

What should scientists do about issues such as GM foods where there is widespread resistance to the technology, clearly arising from very deeply felt convictions, some stemming from concerns that lie outside science?

How society uses science

I believe we must now become much more involved in public debates on how our society uses science and technology to create wealth and a better quality of life. Otherwise, science will become increasingly marginalized. I believe, too, that we scientists must become much more sophisticated about our relationship with the public, our response to pressure groups, and especially our dealings with the media in the following ways.

First, we should not assume that when we explain what we are doing, the public will always agree with us. Scientists and the public often work with different value systems: for example, not every scientific discovery is accepted as being good for society, to take nuclear power and the value of animal experiments as just two instances.

Second, we need to be much clearer about the complex area of risk and risk perception. This is not objective: different risks are weighted differently by different groups in society (the contrast between US and European attitudes to GM technology is a case in point). Risk decisions use judgement as well as evidence; policy decisions often have to be made before all the relevant facts are available, for example where there is a public-health concern, as in the cases of BSE and, currently, mobile phones. Changes in societal attitudes have undermined the role of the 'trusted scientist' as a respected source of advice. Others have to be brought into the decision-making, with the resultant problem that everyone wants to be consulted — but not everyone can be involved.

Third, how should scientists handle the kind of storm that erupted last year over Arpad Pusztai's claims, subsequently found to have no scientifically credible foundation (see *Nature* **401**, 731; 1999), of damage caused to rats' guts by GM potatoes? Who should respond to comments by media interviewers such as: "So the British public

are to be treated as unsuspecting guinea-pigs once again"? Not retailers, who have been quick to sacrifice GM food products rather than lose market share. Not individual companies: Monsanto's brief press campaign was unsuccessful. The public debate is being left, by default, in the hands of scientists.

There are signs that we are learning how to form a constructive dialogue with the public. A conference in Edinburgh in February included representatives from at least 14 developing countries and all major 'green' groups. It achieved a crystallization into points of agreement, points of disagreement, and points where knowledge is currently lacking. There is surprisingly widespread acceptance that the issues are not all black

We should not assume the public will always agree, even if we do explain what we're doing.

and white, but need to be addressed on a case-by-case basis (see <http://www.oecd.org/subject/biotech/edinburgh.htm>).

A subsequent e-mail conference, sponsored by the UN's Food and Agriculture Organization (FAO), on food production in developing countries, is open to all at biotech-admin@fao.org and has attracted 130 submissions from 19 countries, most in the developing world. Their overwhelming message, as in Edinburgh, is that the developing world needs and wants GM technology, but it must involve appropriate partnerships between the public and the private sectors.

The agrifood industries have now (not before time, given their enormous resources) taken two initiatives. CropGen, a UK communications initiative comprised of a panel "from a variety of disciplines including the biological sciences and consumer affairs", is claimed to be independent from its sponsors, who "have all signed an undertaking not to veto any of the scientific positions or influence the activities of CropGen". A similar initiative, sponsored by the Council for Biotechnology in the United States, is funded by seven agrifood companies and the Biotechnology Industry Organization, with a \$50 million budget over three to five years. Whether its advice will be considered 'objective' by the public remains to be seen.

A more hopeful source is the European Union, where Commissioner Philippe Busquin has just set up a biosciences high-level group to give advice on the "scientific aspects of social controversies about biosciences and biotechnology" (see *Nature* **405**, 54; 2000). At the first meeting last month I was impressed by the commissioner's determination to bring in scientific advice at the highest level. We should do all we can to help this new initiative.

Does GM technology need a permanent forum to assess its science and the social implications, as suggested at Edinburgh (see *Nature* **404**, 112; 2000)? One analogy, the Intergovernmental Panel on Climate Change, has been outstandingly successful, despite some criticism. But for GM foods, what is viable? How can all the stakeholders be involved, and under whose auspices: the UN, the FAO, the WHO? Any such body must be open, transparent and inclusive. Is there enough international political will to make this work? I remain to be convinced.

For scientists in academic institutions, listening to the public — as advocated by the House of Lords (see <http://www.publications.parliament.uk/pa/ld199900/ldselect/ldscitech/38/3801.htm> and *Nature* **404**, 211; 2000) — is made easier by two e-mail information networks, one run by the Royal Society in London and the other by the University of Bern, to help those who are approached by the media.

Over the past two years the Royal Society has become more proactive, to help meet its objective of ensuring that policy decisions are based on sound science. It is working with six academies — the US National Academy of Sciences and five from less developed countries — to produce a statement on GM plants.

Still trusted by the public

The traditional inclination of academic researchers has been to comment on, but not to become directly involved in, public controversy. That is no longer enough.

Recent surveys show that scientists, as a group, still have the public's trust. Surely, therefore, we have a duty to help the public make decisions, particularly in areas where science is the subject of campaigns by environmentalist and other special-interest groups. Let's get to it! ■

Derek Burke is at 13 Pretoria Road, Cambridge CB4 1HD, UK. He is the former chairman of the UK advisory committee on novel foods and processes, and former vice-chancellor of the University of East Anglia.
e-mail: dcb27@cam.ac.uk

The house that Jim built

Writings from a scientist who has never rested on his Nobel laurels.

A Passion for DNA: Genes, Genomes and Society

by James D. Watson, edited by Walter Gratzer

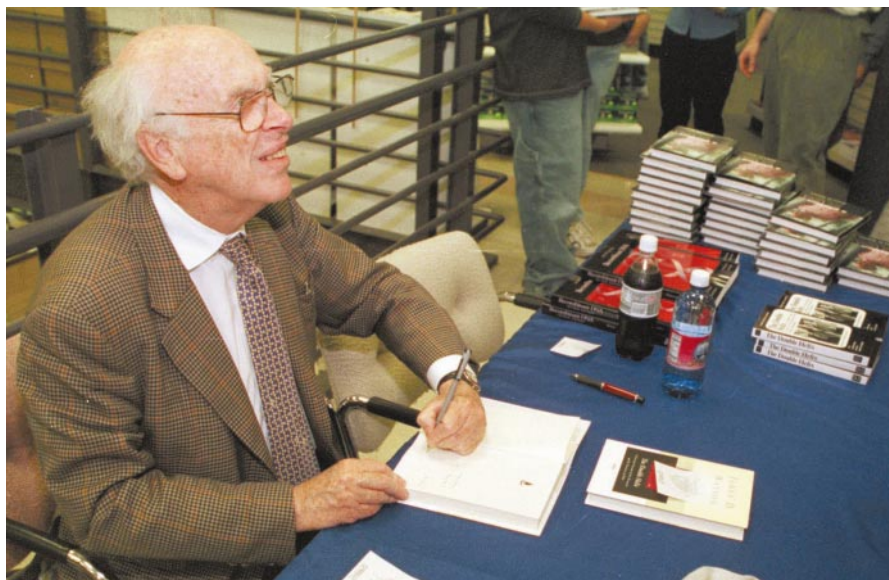
Oxford University Press: 2000. 250 pp.
£18.99, \$25

Sydney Brenner

We need to remind ourselves that *The Double Helix*, Jim Watson's unconventional story of the great discovery of the structure of DNA, was published as late as 1968, some 15 years after the event and six years after the Nobel prize. Jim was 40, and soon to leave Harvard to become, as he puts it, the manager of science at Cold Spring Harbor. By this time, molecular biology had also become an established academic subject, with journals and professors. Undergraduates were memorizing the genetic code in case they were asked a question on it in their exam papers. The heroic period was over, and many of the small band of radical evangelists, who had created the revolution in biology in the wake of the double helix, had become respected pillars of scientific society and the new scientific establishment.

As we learn from these essays, Jim never rested on his laurels, nor did he ever relax his vigilance, finding the enemies of science and dealing with their pomposity and inanity. He also assumed the unlikely role of statesman, and the international human genome sequencing programme owes its existence to Jim's skill in converting most of its early opponents and to his shrewd understanding of Washington politicians. He has used his considerable reputation in the service of science and it is this that has given him the right to make many of the outrageous statements (all carefully calculated) about the war on cancer, the dangers of recombinant DNA, cloning people and other follies of modern science.

Several essays in the book deal with the controversies about recombinant DNA technology, or genetic engineering, as it was once called. In 1973, together with several other scientists, Jim signed the famous statement calling for a moratorium on recombinant DNA research. The effect of this statement in England was for the Medical Research Council to ban this work in its laboratories. The ban was totally effective, mainly because most people did not appreciate that there is a difference between chastity and impotence. Jim recalls how, at the meeting at Asilomar in February 1974, called to discuss the matter, he stood up and said that it was a lot of nonsense, that there could be no rational guidelines and so we might as well all go home.



Further writings: Watson at a signing session for *A Passion for DNA*.

He was right, of course, and although he does admit that calling for a moratorium was a mistake, the only way to end that moratorium was through guidelines. When these finally appeared they had no rational basis and condemned scientists to a series of empty ritual exercises to satisfy the requirements. I can well remember the cry of horror that greeted my remark that I would prefer to work with the Lassa fever virus cloned in *E. coli* than with the virus itself.

We still live today with the confused heritage of those times, with many people seemingly more anxious to discuss the ethical, legal, political and social aspects of the conjectural consequences of new technology than applying themselves to the real problems of mankind such as the HIV epidemic in Africa.

Jim created one of the world's great centres for biological research in Cold Spring Harbor. When he took it over more than 30 years ago it was in serious decline. He introduced new programmes of research and, together with a group of younger scientists, was able to obtain large grants to support the new initiative. He also proved adept at raising money for new buildings and for the restoration of the grounds. Together with his wife Elizabeth he has created a jewel, and old hands such as myself, who can remember the old CSH, find it a great pleasure to visit.

Jim also wrote a wonderful book, *The Molecular Biology of the Gene* (Benjamin Cummings, 1987) — the serious *Double Helix* — which explained the new science and was completely different from every other textbook written at the time. In fact, much of

the modern style in presenting biology comes from that book.

We return — inevitably — to *The Double Helix*. It is to Jim's lasting credit that he retained the immediacy of the discovery and preserved his younger self in the memoir. As he mentions in one of the autobiographical essays, he wanted it to be seen as a work of literature and not as a scholarly tome on the history and philosophy of science. Its style is novelistic, and he confirms here that he thought of it originally as a novel, a work, perhaps, of 'faction' as it came to be applied to books such as Truman Capote's *In Cold Blood*. Perhaps it paints a somewhat exaggeratedly idyllic picture of Cambridge and England, with a little of the veneer of the sixties added to it. After all, in 1953, food was still rationed in England, as I well remember when Jim came to visit us in Oxford in July 1953, and ate most of the cakes at my son Jonathan's seventh birthday party.

When the book appeared, I wrote a review of it for myself which was never published. I compared it to another book of that period, J. D. Salinger's novel *The Catcher in the Rye*, which I read in America in the summer of 1954. This describes the problems of a 16-year-old adolescent, Holden Caulfield, as he struggles to come to terms with the realities of the adult world. He finds most of the people in that world pompous and ridiculous, except for an elder brother, D.B., who lives in Hollywood and has cars and girls. When asked what he'd like to become, he says he couldn't be a scientist, because he is no good at science, and law does not appeal to him. Having mistaken the poem by Robert

book reviews

Burns as reading “If a body *catch* a body, coming through the rye”, he says he’d like to be a catcher in the rye, the only tall person to catch the thousands of little kids playing in the field of rye and saving them from running off the cliff at the edge of the field. He knows it’s crazy, but that’s what he’d like to be. ■

Sydney Brenner is at the Molecular Sciences Institute, Berkeley, California 94704, USA.

Mozart, McCartney and sexual selection

The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature

by Geoffrey Miller

Heinemann/Doubleday: 2000. 528 pp. £20/\$27.50

Margaret A. Boden

This book is an exhilarating example of joined-up thinking — not to mention elegantly joined-up writing. Indeed, it’s so joined-up that no one reviewer could do it justice. It ranges from the origins of *Homo sapiens*, through the details of human sexual anatomy, to the ubiquity of humour, the glories of Mozart and the romance of “flowers that fade, candles that burn, overpriced dinners, and walks on exotic beaches”. And it seeks to explain all these things in evolutionary terms.

But how? As Geoffrey Miller remarks, romantic presents don’t increase a woman’s survival prospects as much as they reduce a man’s bank account. Why does a man bother? And why is a woman amenable in the first place? As for Mozart, how could evolution account for the hard work even he had to put into his composing, or for the international music industry that rests on Mozart — and the Beatles?

And what explains other cultural and aesthetic products — science, dance, craftwork, politics, religion ...? One might say that these things feature in human cultures because they bring pleasure, of one sort or another. But pleasure, Miller repeatedly insists, is what evolutionary theory must explain.

It does this primarily by way of Darwin’s second evolutionary principle: not natural selection, but sexual selection. Despite obvious examples such as the peacock’s tail, says Miller, modern biologists and evolutionary psychologists — not to mention laymen — commonly ignore this principle when they apply evolutionary thinking to psychological and cultural phenomena. He discusses countless cases of anatomical, psychological and cultural features whose explanation involves sexual selection.

That explanatory reference, however, may

be highly indirect. Miller is nothing if not subtle. He’s well aware that a detailed evolutionary explanation of Mozart’s 40th or McCartney’s *Yesterday* is no more likely than an explanation of why this hummingbird’s tail is green and ten centimetres long, whereas that one’s is red and much shorter. Indeed, it’s even less likely. We can, perhaps, hope to find out just which genes, and which developmental conditions, influence the colour and length of hummingbirds’ tails, and which biochemical mechanisms are involved. But to explain how Mozart wrote his 40th, or why *Yesterday* is now considered a classic, would require several even more difficult achievements.

We would need a clear statement of the aesthetic principles involved, and an evolutionary explanation of why those features give us pleasure. We’d need a detailed cultural-historical understanding of the musical styles of the times.

We would need knowledge of Mozart and McCartney as individual artists, with their own personal ‘signatures’, and of any chance psychological associations involved. Not least, we’d need an understanding of how human creativity is motivated, how it functions at the cognitive level, and why it evolved in the first place. Don’t hold your breath.

But that’s not to say that none of these questions can be fruitfully approached from an evolutionary viewpoint. Miller has much to say about the origins of creativity in general, and also of some particular aesthetic principles. He regards human creativity as an example of “protean” (unpredictable) behaviour. Unpredictability has co-evolved in predator–prey behaviour, for instance. But that is explicable by Darwin’s first principle: natural selection. How can protean behaviour — and protean thinking — be favoured by sexual selection?

To cut Miller’s long, and carefully argued, story short, increased brain-size and behavioural flexibility formed a positive-feedback loop, because of runaway sexual selection. Within limits, surprise is not only attention-getting but motivating — which is to say, pleasurable. To be surprised, and to be bored, implies a capacity for recognition, or even prediction. In short, both surprise and boredom require brain cells.

Similarly, the anticipation and pre-emption of boredom require brain cells. The more complex the expected patterns, and the more subtle the surprises, the more brain-power is required to generate and appreciate them. So the females’ attention is captured by the more surprising males, but the females themselves need intelligence to be surprised in the first place. And the males, naturally (*sic*), appreciate being appreciated: a female too dim to recognize one’s adventurousness is not a good prospect for mating. So male–female co-evolution ratchets up the size of the cerebral cortex until the anatomy of the birth



Vistas of love

Love & Desire: Photoworks by William A. Ewing (Thames & Hudson/Chronicle Books, £16.95/\$35) is a collection of photographs showing love, affection and desire in their myriad forms.

canal calls a halt. In short, the large brains and complex intelligence of *Homo sapiens* aren’t due to natural selection (how many brain cells do you need to pick a berry, or catch a buffalo?), they’re due rather to sexual selection — but they come in handy later for survival purposes.

Cultural activities such as craft, music, decoration and dance provide not only pleasurable surprises (remember Robert Herrick’s poem, “A sweet disorder in the dress ...”), but also fitness indicators. These may be more or less direct. Someone who can join in energetic and precisely patterned dancing for hours clearly has physical strength and accomplished body control, and may be an acceptable mate. (Compare the male peacock, strong enough to grow and carry that cumbersome tail.)

Or consider symmetry. It’s well known that facial and bodily symmetry contribute to judgements of sexual attractiveness or physical beauty. Miller suggests that the appreciation of symmetry in cultural artefacts (not only pottery and jewellery, but ritual practices and mathematics) utilizes the same cognitive mechanisms, and has similarly positive motivating effects. Bodily symmetry is preferred because it is an indicator of biological fitness: symmetry is hard for the genes to achieve. And it’s hard for human craftsmen to achieve, too. Someone capable of the physical effort and self-control needed to make a perfectly symmetrical flint axe-head is fitter (in the Darwinian sense) than someone who is not, and a better prospect for mating. As Miller points out, a high proportion of superbly crafted prehistoric axe-heads are

not only beautifully polished, sharp and symmetrical, but — as electron microscopy shows — were never actually used. So why were they made? His answer is that they were valued, as we say, for their own sake — in other words, for their ability to give us pleasure, which in turn rests on their value as fitness indicators of the men who made them.

Miller has many stimulating and unexpected things to say about these, and other, cultural matters. He's also good at jokes. His book will excite many people, and infuriate others — especially if they don't read it carefully. *The Mating Mind* is as far from naive biological determinism as it could be, without losing sight of our evolutionary heritage altogether. Reading this book enables a rich confusion of apparently unrelated details to fall into place — and isn't that what science is all about? ■

Margaret A. Boden is in the School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QN, UK.

..... Statistical detective

Statistics on the Table: The History of Statistical Concepts and Methods

by Stephen M. Stigler

Harvard University Press: 1999. 448 pp.
\$27.95, \$45

Ian Hacking

Stephen Stigler's previous book *The History of Statistics: The Measurement of Uncertainty before 1900* was a rare but splendid example of the history of a science written by one of its practitioners. It also conveyed a wry wit and insatiable curiosity that were not, however, allowed to distract from the main story. Stigler has also written occasional articles over the years, expressing his detective's delight in uncovering obscure but charming facts. In *Statistics on the Table* he has revised these pieces, which, although organized to develop a number of systematic lines of thought, are often so interesting individually that this could serve as a bedside book.

The title comes from a private letter by Karl Pearson, who wrote that, since you can draw any number of conclusions — sound and unsound — from statistical data, you should always put your data on the table and show how you analyse them before you state your inferences. Stigler shows how the desire to present data in compact, unbiased and accessible ways motivated the founding figures of his discipline. He also applies the motto to himself, setting out the data that bear on his own historical problems and telling how he came by them.

His focus is on figures who now seem minor; some you would never encounter

without Stigler as guide. Even when he turns to the famous, he directs us to the margins of their careers. Yet tucked away at the end of many of these pieces is a gently stated but powerful methodological moral for our times.

The essays are arranged in five sections. "Statistics and Social Sciences" begins with Pearson's relationships with the Cambridge economists, including the greatest: Marshall and Keynes. There are two chapters on Jevons and a very fine essay on Edgeworth, important enough in their day, but now almost unknown. After a bicentennial homage to Quetelet, we proceed to Galtonian notions. Stigler guides us through the traps of regression, with special reference to a bizarre series of tables published in 1933 by one Horace Secrist, who proved that American business was (necessarily, as shown by the mathematical theory of regression, and, in fact, as established by observation) regressing towards mediocrity. What seems like anecdote is, in fact, a veiled denunciation of the number-crunchers who employ vast software programs with no idea of the ideas they deploy, and no comprehension of the numbers that result.

As we progress through the book Stigler the detective becomes more and more in evidence. A bookseller's list piques his curiosity with a 1695 tract on *The Art of Curing Diseases by Mathematics*, written by Sir Edward Eizatz to denounce 'A.P.'. This leads us to Edinburgh poet and physician Archibald Pitcairne (1652–1713), who opposed the unscientific medical practices of his day with a new theory of medicine based on mathematics. Any morals to draw here? This spat, by no means foolish in its details, is an example of what happens when mathematics tries to intrude into fields where there are already practitioners.

In his section on naming, we are presented with Stigler's Law of Eponymy. The law states that (on the Principle of Humility) no law, theorem, principle or phenomenon is named after the person who discovered it. Its name is a joke: by self-reference, Stigler cannot have been the first to hit on Stigler's Law and, sure enough, he gives the credit to the sociologist of science Robert K. Merton. Humility triumphs, after all.

Various figures cross this section — Gauss, Euler, Pearson — but "Who discovered Bayes's Theorem?" conveys the flavour. Richard Price published Bayes's most famous paper in 1765, four years after Bayes's death, but Stigler's Law says there must be a forgotten precursor. And yes, there is a mention of what sounds very much like Bayes's ideas in work published in 1749 by David Hartley, founder of associationist psychology. Hartley says of the Bayes-like result that it was communicated to him by an "ingenious friend". Who?

Stigler unearths a remarkable forgotten mathematician named Nicholas Saunderson. A good candidate, on data provided. But what would increase the credibility? In order to possess the requisite mathematics, including the binomial expansion, the friend would have to have read Abraham de Moivre, who set out matters clearly in a miscellany published in 1730. So who read de Moivre? Examine the list of subscribers to the first printing: Saunderson is there. From this and other reflections Stigler derives a list of suspects whom he will summon into the parlour — his metaphor. After he has put all his data on the table, Stigler's 'Bayesian [sic] calculation' gives odds of 3 to 1 on Saunderson. This is vintage Stigler. ■

Ian Hacking is in the Department of Philosophy, University of Toronto, 215 Huron Street, Toronto, Ontario M5S 1A1, Canada.

Aventis Prizes for Science Books 2000

The Elegant Universe by heretical theorist Brian Greene (Vintage, £7.99, \$15, pbk) has won this year's Aventis Prizes General Prize — often called the scientific community's Booker Prize.

The prizes were set up in 1988 by the Science Museum and COPUS, the Committee On the Public Understanding of Science, to encourage the writing, publishing and sale of popular science books for non-specialist readers. This year's general prize, worth £10,000 (£15,000), attracted a record 117 entries. Other books on the shortlist are shown below. (See overleaf for coverage of the junior prize.)

The White Death: A History of Tuberculosis

by Thomas Dormandy

Hambledon/New York University Press, £25, \$29.95

A Brief History of the Future: The

Origins of the Internet

by John Naughton

Weidenfeld & Nicholson/Overlook, £18.99/\$27.95

Genome: The Autobiography of a Species in 23 Chapters

by Matt Ridley

Fourth Estate/HarperCollins, £9.99 (pbk)/\$26 (hbk)

Time, Love, Memory: A Great Biologist and his Quest for the Origins of Behaviour

by Jonathan Weiner

Faber & Faber/Knopf, £20/\$27.50

Children of Prometheus: The Accelerating Pace of Human Evolution

by Christopher Wills

Allen Lane/Perseus, £20 (hbk)/\$15 (pbk)

Making sense of a small world

The Aventis Junior Science Book Prize 2000 shortlist

Maxine Clarke

If you need to make a tricky decision quickly, co-opt a nine-year-old. With the unnerving decisiveness of youth, my co-reviewer took less than five minutes to select her winner among the six books on the shortlist for the under-14 category of the annual Aventis science book prize (see Box, below). The adult judges who chose the shortlist handed over to pupils aged 8–14 from 31 UK schools to select a winner. The professional help thus enlisted, from a population not only unaware of the concept of charging by the hour but temperamentally alien to it, coughed up a winner in short order, undoubtedly avoiding the hours of solemn deliberations that adults feel is necessary before making decisions.

One of the finalists, *The Usborne First Encyclopaedia of Our World*, is comprehensible to a four-year old, whereas some of the others deal with concepts fairly challenging even for a secondary-school student, older than 11. The younger child loves being read almost anything by an adult, but the encyclopaedia format is probably not optimal

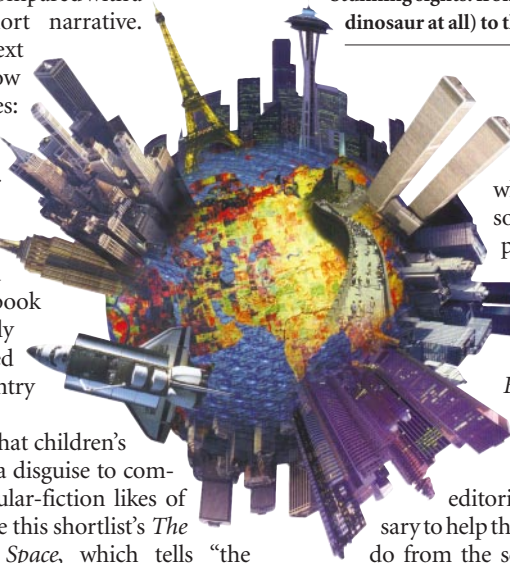
for this age-range compared with a sustained but short narrative. So my advice for next year is to narrow down the categories: the under-7, 8–11 and 12–14 age groups are different species. Older readers

may well find the Usborne book too simplistically presented compared with Kingfisher's entry on a similar topic.

It's a tradition that children's non-fiction needs a disguise to compete with the popular-fiction likes of *Goosebumps*. Hence this shortlist's *The History News in Space*, which tells "the enthralling story of the Russian and American space race this century" as tabloid newspaper journalism in exciting bite-size portions — with boxes explaining concepts such as Newton's third law as well as (disappointingly, spoof) ads for souvenir flight badges and other merchandise. Although this might be rather fun to an adult, it is unclear how the young reader is supposed to distinguish the factual entries from the fictional ones.

Two other titles fall into the teaching-by-stealth category by packaging

Stunning sights: from the plesiosaur (not a dinosaur at all) to the modern world.



themselves as joke books: both are cheap, black-and-white paperbacks and so are instantly appealing to the target age-group. One is *Evolve or Die* (see Box). The other, *Brainwaves in the Bedroom*, purports to be about magic but is really about neuroscience. Far more editorial guidance is necessary to help the reader tell the pseudo from the science, a distinction difficult enough for many adults.

DK Guide to Space and *The Kingfisher Book of Planet Earth* are firmly in the traditional mould, both beautifully illustrated, large-format hardbacks with a new subject on each page. The Kingfisher book is a stunningly presented account of the history of our planet, full of wonderful colour drawings with photographs used relatively sparingly. The text is written with an attractive blend of fact and enthusiasm by Martin Redfern — the opening states "Five billion years ago, there was no planet Earth and no Sun. But the Universe was already in full swing". He continues in this vein with clear, engaging explanations of the history of the planet: volcanoes, earthquakes, geology, origins of life and, finally,

a beguilingly didactic last page entitled "The end of the Earth", stating the four most likely ways in which annihilation will occur. A few pages of useful data and a glossary finish the book off neatly. This one is my favourite, for its accessible style and for getting to grips with topical problems such as anthropogenic climate change and declining biodiversity. Adults buying this book for their children won't be able to stop themselves reading (and learning from) it, too.

DK Guide to Space has big photographs and not many words on topics such as "star birth", "eclipses" and "man on the moon". The text is unambiguously informative: jokes and tricks have no place here. Peter Bond writes simple, clear, even poetic explanations: "A vast frozen ocean completely covers the moon Europa. When sunlight catches its icy face, Europa is a dazzling white." Only one page of data at the back, but this is bang up to date with a table of website addresses for the reader stimulated by the book to find out more. It clearly stimulated the young judges, since this was the one they chose as the winner.

Maxine Clarke is on the editorial staff of *Nature*.

The view from Year 4

Catherine Irving

Evolve or Die is my favourite. It is informative and makes people laugh because of all the silly jokes and pictures. The pictures are not always right, for example beetles are drawn like little men, but this is OK because the words are serious and it is a funny way of making the subject interesting. It asks questions to check you are reading it properly — a good idea.

Brainwaves in the Bedroom is good because it tells you that magic and science have a lot in common. You have to do lots of puzzles and it is fun working out the tricks. This book is trying to tell you how your brain understands things.

First Encyclopaedia of Our World is very scientific because it shows you what real scientists do, and things like where all the planets are in space, and how a storm begins and how people actually find fossils and dinosaur bones. The pictures show you a lot, so if you read the words it helps them make sense.

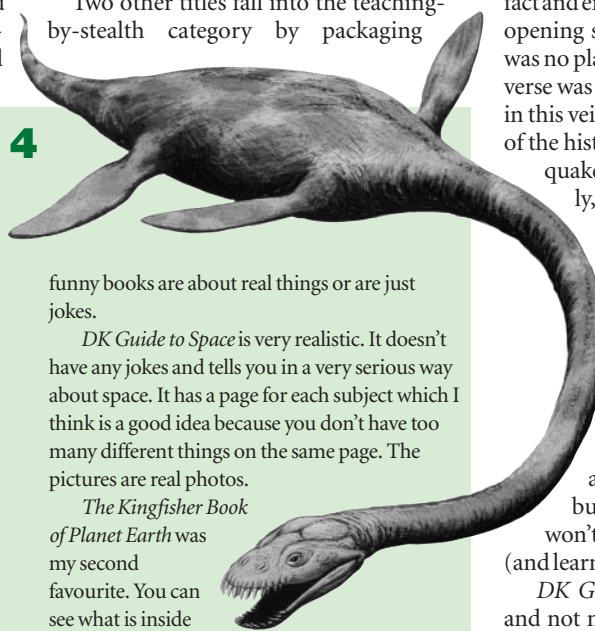
The History News in Space had the very good idea of putting it all as newspaper cuttings. I also thought it was good because it shows you how rockets are made and which famous people discovered things about space. I like the newspaper idea of explaining science because it is realistic and you don't know if the words in the

funny books are about real things or are just jokes.

DK Guide to Space is very realistic. It doesn't have any jokes and tells you in a very serious way about space. It has a page for each subject which I think is a good idea because you don't have too many different things on the same page. The pictures are real photos.

The Kingfisher Book of Planet Earth was my second favourite. You can see what is inside the Earth and it shows you how life began, and has really good pictures like skeletons of old creatures that have died out. I think it is very scientific: it doesn't have any jokes in it. Each page has a good title like "The Real Jurassic Park" that makes you stop and look. Then you look at the pictures and this makes you want to read the words.

Catherine Irving is in Year 4 at St Luke's Church of England Primary School, Kingston on Thames, Surrey, UK.



A double-edged sword

The technical fix for one genetic disorder had unforeseen repercussions.

Diane Paul

In 1960, Robert Guthrie, a microbiologist and physician working in Buffalo, New York, devised a simple way to measure the level of phenylalanine in the blood of children with the genetic disorder phenylketonuria (PKU). Guthrie's interest in the problem was intensely personal — he had a son and a niece who were mentally retarded. The niece's condition was due to PKU, in which a defective liver enzyme results in an inability to metabolize phenylalanine, an essential amino acid found in all dietary proteins.

In such patients, phenylalanine accumulates to toxic levels, often causing severe mental retardation and physical and behavioural abnormalities. Physicians reasoned that at least some children could be helped by a diet from which most of the offending amino acid had been removed. But before Guthrie, there was no practical way to monitor blood phenylalanine levels.

Building on his work on measuring drug levels in cancer patients, Guthrie developed such a test. It involved adding a blood specimen, on a disk of filter paper, to a culture of bacteria unable to metabolize their own phenylalanine. If the blood came from an individual with PKU, the bacteria sprang into life.

Although this was an important achievement, the key to controlling PKU seemed to lie in diagnosing the condition as soon after birth as possible, before the infants' developing brains were damaged. The existing 'wet diaper' test, in which a drop of ferric chloride added to a urine-soaked diaper turned green if the infant had PKU, was unreliable until the age of eight weeks, by which time the infant might have suffered irreversible brain damage. Guthrie's method was much more sensitive, detecting abnormal blood phenylalanine levels by the third day of life. It was also convenient and — partly due to Guthrie's refusal of royalties — inexpensive.

Guthrie's test converged with the commercial release of a PKU formula in 1958. At this time, general frustration over the inability to treat mental retardation was shifting the emphasis from providing the retarded with social and rehabilitative services towards scientific treatment. In 1961, President John F. Kennedy, whose sister had long been institutionalized for mental retardation, announced a major national campaign. In this context, the Guthrie test received enormous publicity and government support. Even before a government-sponsored trial of the test had ended, the state of Massachusetts made it mandatory, and by



Screen test: a simple procedure can identify infants with the genetic disorder phenylketonuria.

the mid-1970s, PKU testing was routine.

The control of PKU generated such excitement because it seemed to offer a template for future successes. PKU is rare, occurring only once in every 11,000–15,000 births in the United States and most of Western Europe. But it was hoped, and often assumed, that dietary therapy would cure many other forms of mental retardation and illness. Alas, this was not to be.

Even in PKU, treatment turned out to be prolonged and arduous. When mass screening began, it was expected that children could resume a normal diet at about age five, when gross brain development was complete. But as evidence grew that ending the diet could result in significant deterioration, scientists became more cautious. It is not easy, however, to follow a 'diet for life'. And even individuals who are treated early and well often have subtle cognitive and behavioural impairment.

Both preparing and eating the PKU diet is burdensome: it requires substitutes for most protein foods, such as bread, rice, pasta, meat, fish, eggs and dairy products. These are supplemented with a special formula containing the other amino acids. Both the low-protein foods and the formula are

expensive. And, of course, many of the social pleasures of eating are lost.

Individuals with PKU do not usually feel sick, so they lack an important incentive to comply with the diet. Compliance is particularly hard for pregnant women, who must drink more of the formula at a time when many are nauseated by it. If they fail to maintain metabolic control, the effects on their offspring can be catastrophic. At high levels, phenylalanine is a teratogen — over 90% of the infants of untreated PKU mothers will be irreversibly retarded. This problem of 'maternal PKU' is an unintended consequence of screening. Before the Guthrie test, few women with PKU bore children. Now, fortunately, many marry and have families. But unless maternal PKU can be controlled, much of the benefit of screening might be erased by the birth of children to untreated women.

PKU has become an exemplar for the promise of genetic medicine. It is true that screening and treatment have allowed many individuals who would have been severely damaged to attend school, hold jobs, and live otherwise normal lives. But the history of PKU also provokes scepticism about simple scientific fixes. 'Solving' the initial problem created the new one of maternal PKU. Moreover, for men and women, coping with PKU has turned out to be a lifelong effort, requiring substantial psychological, financial and social support. The path of scientific prevention did not, after all, obviate the need for social services; if anything, it has increased their importance. ■

Diane Paul is in the Department of Political Science, University of Massachusetts at Boston, 100 Morrissey Boulevard, Boston, Massachusetts 02125. This essay draws on collaborative work with Paul J. Edelson.

Phenylketonuria has become an exemplar for the promise of genetic medicine.

Catching crumbs from the table

In the face of metahuman science, humans have become metascientists.

Ted Chiang

It has been 25 years since a report of original research was last submitted to our editors for publication, making this an appropriate time to revisit the question that was so widely debated then: what is the role of human scientists in an age when the frontiers of scientific inquiry have moved beyond the comprehensibility of humans?

No doubt many of our subscribers remember reading papers whose authors were the first individuals ever to obtain the results they described. But as metahumans began to dominate experimental research, they increasingly made their findings available only via DNT (digital neural transfer), leaving journals to publish second-hand accounts translated into human language.

Without DNT, humans could not fully grasp earlier developments nor effectively utilize the new tools needed to conduct research, while metahumans continued to improve DNT and rely on it even more. Journals for human audiences were reduced to vehicles of popularization, and poor ones at that, as even the most brilliant humans found themselves puzzled by translations of the latest findings.

No one denies the many benefits of metahuman science, but one of its costs to human researchers was the realization that they would probably never make an original contribution to science again. Some left the field altogether, but those who stayed shifted their attentions away from original research and toward hermeneutics: interpreting the scientific work of metahumans.

Textual hermeneutics became popular first, since there were already terabytes of metahuman publications whose translations, although cryptic, were presumably not entirely inaccurate. Deciphering these texts bears little resemblance to the task performed by traditional palaeographers, but progress continues: recent experiments have validated the Humphries decipherment of decade-old publications on histocompatibility genetics.

The availability of devices based on metahuman science gave rise to artefact hermeneutics. Scientists began attempting to 'reverse engineer' these artefacts, their goal being not to manufacture competing products, but simply to understand the physical principles underlying their operation. The most common technique is the crystallographic analysis of nanoware appli-

ances, which frequently provides us with new insights into mechasynthesis.

The newest and by far the most speculative mode of inquiry is remote sensing of metahuman research facilities. A recent target of investigation is the ExaCollider recently installed beneath the Gobi Desert, whose puzzling neutrino signature has been the subject of much controversy. (The portable neutrino detector is, of course, another metahuman artefact whose operating principles remain elusive.)

The question is, are these worthwhile undertakings for scientists? Some call them a waste of time, likening them to a Native American research effort into bronze smelting when steel tools of European manufacture are readily available. This comparison might be more apt if humans were in competition with metahumans, but in today's economy of abundance there is no evidence of such competition. In fact, it is important to recognize that — unlike most previous low-technology cultures confronted with a high-technology one — humans are in no danger of assimilation or extinction.

There is still no way to augment a human brain into a metahuman one; the Sugimoto gene therapy must be performed before the embryo begins neurogenesis in order for a brain to be compatible with DNT. This lack of an assimilation mechanism means that human parents of a metahuman child face a difficult choice: to allow their child DNT interaction with metahuman culture, and watch him or her grow incomprehensible to them; or else restrict access to DNT during the child's formative years, which to a metahuman is deprivation like that suffered by Kaspar Hauser. It is not surprising that the percentage of human parents choosing the Sugimoto gene therapy for their children has dropped almost to zero in recent years.

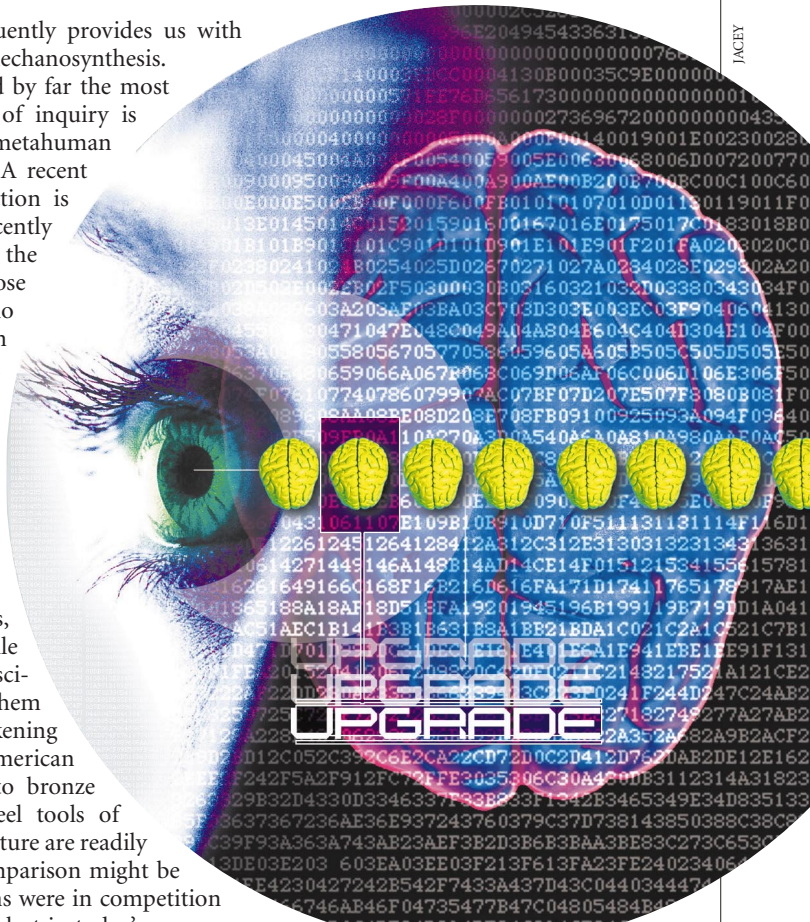
As a result, human culture is likely to survive well into the future, and the scientific tradition is a vital part of that culture. Hermeneutics is a legitimate method of sci-

entific inquiry and increases the body of human knowledge just as original research did. Moreover, human researchers may discern applications overlooked by metahumans, whose advantages tend to make them unaware of our concerns.

For example, imagine if research offered hope of a different intelligence-enhancing therapy, one that would allow individuals to gradually 'upgrade' their minds to a level equivalent to that of a metahuman. Such a therapy would offer a bridge across what has become the greatest cultural divide in our species' history, yet it might not even occur to metahumans to explore it; that possibility alone justifies the continuation of human research.

We need not be intimidated by the accomplishments of metahuman science. We should always remember that the technologies that made metahumans possible were originally invented by humans, and they were no smarter than we.

Ted Chiang is an occasional writer of science fiction. His latest story can be found in the anthology *Vanishing Acts*, published by Tor Books.



Pax Argentina

David C. Queller

Colonizing species often lose genetic variation. Usually this has harmful effects, but in Argentine ants it has led to loss of clan warfare and formation of supercolonies that overwhelm native species.

In the first and second centuries, most of the Mediterranean region flourished under the Pax Romana. Wars with foreign enemies continued at the frontiers, but the peaceful core of the empire gave border provinces the advantage of friendly forces at their flanks and rear. A paper by Tsutsui and colleagues, just published in *Proceedings of the National Academy of Sciences*¹, shows how a biological version of this system has been reinvented by the Argentine ant, *Linepithema humile*. The study yields some fascinating insights into the causes of both ecological and evolutionary success.

The Argentine ant is a superb invader of non-native habitats, particularly, as it happens, around the Mediterranean and areas with similar climates such as California. It achieves high population densities, often controlling huge areas to the near exclusion of other ants². We rarely understand why invading species succeed, although a common advantage is that they leave their predators, parasites and pathogens behind. Although this explanation could apply to Argentine ants, it seems that the most serious enemies left behind were the warring clans of its own species.

In their native Argentina, ants from different colonies fight (Fig. 1), as most ants do, and the species is not ecologically dominant³. Its dominance in non-native areas seems to be associated with a very unusual trait. There, the nests are not distinct colonies. Instead, they form a vast supercolony within which there is little aggression and extensive interchange of both workers and queens. Such ants are called unicolonial because a whole population effectively becomes one colony⁴.

The advantages of peace are shown by lab experiments pairing Argentine ant colonies that were either mutually tolerant or mutually aggressive. Tolerant pairs had lower

mortality, and higher feeding and growth rates⁵. Earlier field studies in California^{2,6} showed that the high population densities associated with tolerance help Argentine ants to outcompete native ants, towards which they are anything but tolerant. The Argentine ants win, not because of the individual prowess of the fighter-workers, which are 2–3 mm in length, but because of their ability to rapidly recruit legions of troops from their network of nests. So peace with flanking nests generates advantages in competition with other species.

How was this Pax Argentina attained? The Romans achieved their peace in part by extending citizenship to their former foes. Tsutsui *et al.*¹ have now shown that the Argentine ants introduced into California and elsewhere took a similar step, albeit inadvertently. Colony membership in social insects — their citizenship — is based on kinship. Usually colony boundaries are fiercely guarded because the only way that non-reproductive workers can pass on their genes is to ensure that their work benefits a closely related reproductive queen. So how a species evolves to ignore colony boundaries is a real puzzle.

Tsutsui *et al.* show that colonies in native Argentina behave as expected, in that aggression depends on the degree of genetic difference, as assessed by seven microsatellite loci in nuclear DNA (these are useful markers because of their high variation and because they are probably neutral in terms of natural selection). Closely related colonies are tolerated; others are not. The population introduced into California has, not surprisingly, passed through a genetic bottleneck — that is, a temporarily small population in which

genes are lost through chance effects known as genetic drift. In this case the ants lost two-thirds of the variability (heterozygosity) at the microsatellite loci. So all of the ants are genetically alike, and those applying the old similarity-tolerance rule could be fooled into accepting everyone as kin.

Support for this view came when Tsutsui *et al.* found a small number of mutually aggressive nest pairs in the Californian population: the relationship between genetic similarity and degree of aggression was the same as in Argentina. So, paradoxically, the ecological success of the introduced populations stems not from adaptation but from the loss of an adaptation — colony recognition — due to genetic drift.

This is not the first time that life has crafted new cooperative units out of formerly independent ones. Many of the major transitions of evolution take this form⁷. Independent molecular replicators became compartmentalized into cells and linked into chromosomes; free-living bacteria became organelles in more complex eukaryotic cells; and cells united in multicellular organisms. A few of these organisms, notably the social insects, evolved colonies that are so cooperative that they are sometimes called superorganisms. In merging colonies to build a super-superorganism, the Argentine ant is one of a few social insects that adds another Russian doll to the series.

Is this the next major evolutionary transition? Do we face a future dominated by unicolonial ants? Some homeowners may feel that this future is already here, but the long-term success of unicolonial species seems doubtful. Although unicoloniality has evolved several times in ants, it does not seem to give rise to large and successful branches of the ant tree⁴. There is a good reason

for this, one that echoes a reason cited by some scholars



Figure 1 In combat — Argentine ants fighting. Fighting is common between colonies in the species' native Argentina, but rare in populations introduced elsewhere. As Tsutsui *et al.*¹ show in studies in California, such peaceful behaviour seems to stem from errors of kin recognition. Paradoxically, these errors, which result in peaceful cooperation between nests, give the species a powerful advantage in battles with competing species.

MARC S. DANTZKER

for the fall of the Roman Empire — a decay of civic virtue.

Without kin discrimination, relatedness in supercolonies becomes indistinguishable from zero⁸, which has two consequences. First, workers that can no longer aid close relatives may evolve more selfish strategies. It may not be coincidental that the best example of ultraselfish 'greenbeard genes' comes from another ant with unicolonial characteristics⁹. Here, workers possessing one allele at a genetic locus selectively murder queens lacking that allele.

Whether or not workers evolve selfish strategies, the second, larger problem is that they cannot improve or even sustain their cooperative behaviour¹⁰. Without relatedness, adaptive modifications of cooperative worker behaviour cannot be favoured and maladaptive ones cannot be disfavoured. Random drift will become important again, this time because of the absence of any opposing force rather than a small population size. The ants are like a casino gambler who is lucky once but cannot quit: chance got them their stake, but over the long run it can only lead to ruin.

Either way, Argentine ants may evolve to control themselves, with their Pax Argentini-

ca falling under an assault of unbridled nepotism, or with their society suffering the lingering death of decay by drift. Of course, either process could take a very long time. As to the question of controlling the ants in California, Tsutsui *et al.*¹ suggest a counterintuitive strategy that might give the ants a push down the nepotistic path. If the main enemies the ants left behind are themselves, perhaps we should introduce more of them. The resulting increase in genetic variation could re-establish the lost kin discrimination. If so, the ants can return to battling among themselves, giving native species a fighting chance. ■

David C. Queller is in the Department of Ecology and Evolutionary Biology, Rice University, PO Box 1892, Houston, Texas 77251-1892, USA.
e-mail: queller@rice.edu

1. Tsutsui, N. D. *et al.* *Proc. Natl Acad. Sci. USA* **97**, 5948–5953 (2000).
2. Human, K. & Gordon, D. M. *Oecologia* **105**, 405–412 (1996).
3. Suarez, A. V., Tsutsui, N. D., Holway, D. A. & Case, T. J. *Biol. Invasions* **1**, 43–53 (1999).
4. Bourke, A. F. G. & Franks, N. R. *Social Evolution in Ants* (Princeton Univ. Press, 1995).
5. Suarez, A. V., Suarez, A. V. & Case, T. J. *Science* **282**, 949–952 (1998).
6. Holway, D. A. *Oecologia* **116**, 252–258 (1998).
7. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
8. Krieger, M. J. B. & Keller, L. *Molec. Ecol.* **9**, 119–126 (2000).
9. Keller, L. & Ross, K. R. *Nature* **394**, 573–575 (1998).
10. Queller, D. C. & Strassmann, J. E. *Bioscience* **48**, 165–174 (1998).

that internal noise can lead to premature switching between stable states³ and irregular oscillations^{4,5} in synthetic gene-regulatory networks. It has been suggested that the inclusion of feedback in gene-regulatory circuits could improve their robustness to internal noise^{3–6}. Until now, that suggestion has remained unverified by direct experimentation.

The steady state of a noisy system can be characterized as a competition between stabilizing system dynamics and destabilizing random fluctuations. The addition of negative feedback counteracts the internal noise by enhancing the stabilizing action of the system dynamics. This effect can be visualized using a simple physical analogy. A ball (representing the level of gene expression) rolling in a salad bowl (representing the gene network) will ultimately come to rest at the bottom of the bowl. Shaking the bowl (the internal noise) will push the ball away from the bottom, while the shape of the bowl will push the ball back towards the bottom. Adding negative feedback to a gene network is equivalent to making the sides of the salad bowl steeper.

Becskei and Serrano¹ use linear stability analysis of a mathematical model to show that negative feedback can double the stability of gene expression in a simple gene network. They confirm this effect with simulations that show a reduction in the variability of gene expression. To test their predictions experimentally, Becskei and Serrano construct a synthetic gene network in the bacterium *Escherichia coli*. The network includes a TetR (tetracycline repressor)-regulated promoter region⁷ placed upstream of the gene encoding the tetracycline repressor itself (Fig. 1). This system constitutes a negative feedback loop, because the promoter is inhibited by the repressor whose expression it drives. To measure the output of the system, Becskei and Serrano fuse the

Gene regulation

Neutralizing noise in gene networks

Timothy S. Gardner and James J. Collins

The thousands of chemical reactions that sustain living cells involve interactions between individual molecules. Such interactions are stochastic, happening discretely and randomly. This so-called noise is inherent in every molecular event that takes place in a cell. In particular, it can cause sizeable, random fluctuations in the concentrations of expressed proteins and RNA. But often the events in an individual cell, or even a whole organism, are not at all unpredictable. The gestation of multicellular organisms, for instance, follows a pattern and time course so predictable it could be used to set a watch. And all too familiar to the frequent traveller is the circadian rhythm, which is driven by a molecular clock with a precise period of oscillation. For years, researchers have wondered how biology achieves such predictability despite its inherent noisiness.

On page 590 of this issue¹, Becskei and Serrano provide insight into a possible mechanism by which cells may accomplish this impressive task. They show, using a synthetic gene circuit, that negative feedback can dramatically reduce variability in gene expression.

Earlier modelling studies² showed that noise in gene expression could lead to qualitative differences in a cell's phenotype if the expressed genes act as inputs to downstream regulatory thresholds. Recent experiments — including our own³ — have also shown

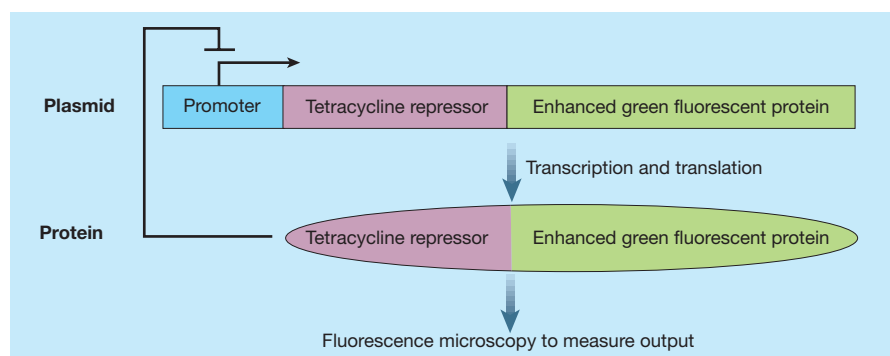


Figure 1 Controlling noise by negative feedback. The synthetic gene network used by Becskei and Serrano¹ includes a tetracycline-repressor-regulated gene promoter to direct transcription of the repressor gene itself. The repressor protein then inhibits transcription from the promoter. The resulting negative feedback stabilizes expression of the repressor around a particular steady-state level. If the repressor exceeds this level, it will strongly inhibit its own synthesis, and repressor levels will decline. Conversely, when the repressor falls below the steady-state level, it will reduce the inhibition and repressor synthesis will increase. The gene encoding enhanced green fluorescent protein is also fused to the gene construct, allowing quantification of the output of the network.

gene encoding enhanced green fluorescent protein to the repressor gene, and quantify expression of the resulting protein by fluorescence microscopy. They then compare this feedback network with control networks that contain the same promoter but lack negative feedback.

Their results show a more than threefold decrease in the variability of gene expression in the feedback network, compared with the control networks. They also show that the improved stability in the feedback network can be reduced incrementally by applying increasing concentrations of anhydrotetracycline — a chemical inhibitor of TetR.

In past theoretical studies of gene-regulatory networks, researchers have typically adopted a reverse-engineering approach, in which principles and mechanisms of function are extrapolated from data about naturally occurring gene networks. A few well-defined natural gene circuits, such as those involved in bacterial chemotaxis⁸ and in the bacteriophage lambda⁹, have emerged as models for theoretical study. But the magnitude and complexity of natural gene networks have, in general, limited researchers' ability to verify their predictions experimentally.

The work of Becskei and Serrano, on the other hand, exemplifies a forward-engineering approach to the study of gene expression. This approach, also adopted in refs 3 and 4, makes use of synthetic gene networks to attain more complete control of the system parameters. Precise perturbations can be introduced to the system without interference to, or from, ancillary processes in the cell. Thus, synthetic gene networks allow

the accurate comparison of theoretical and experimental results, and can, in theory, quickly reveal possible principles and motifs of cellular regulatory processes.

Synthetic gene networks are rapidly emerging as a valuable tool for the identification, analysis and understanding of cell regulatory processes. But it remains to be shown whether observations of a synthetic gene network apply in the context of natural networks. In the case of Becskei and Serrano's work¹, it is still not clear whether *E. coli* or other organisms do exploit negative feedback to stabilize gene expression. When this question is answered, it may not only solve the mystery of biology's extraordinary precision, but may also reveal the full value of the forward-engineering approach to the study of biological systems.

Timothy S. Gardner is at Cellicon Biotechnologies, 12670 Torrey Bluff Drive, Suite 257, San Diego, California 92130, USA.

e-mail: tgardner@cellicon.com

James J. Collins is at the Center for BioDynamics and Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA.

e-mail: jcollins@bu.edu

1. Becskei, A. & Serrano, L. *Nature* **405**, 590–593 (2000).
2. McAdams, H. H. & Arkin, A. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
3. Gardner, T. S., Cantor, C. R. & Collins, J. J. *Nature* **403**, 339–342 (2000).
4. Elowitz, M. B. & Leibler, S. *Nature* **403**, 335–338 (2000).
5. Barkai, N. & Leibler, S. *Nature* **403**, 267–268 (2000).
6. Ko, M. S., Nakauchi, H. & Takahashi, N. *EMBO J.* **9**, 2835–2842 (1990).
7. Lutz, R. & Bujard, H. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
8. Barkai, N. & Leibler, S. *Nature* **387**, 913–917 (1997).
9. Arkin, A., Ross, J. & McAdams, H. H. *Genetics* **149**, 1633–1648 (1998).

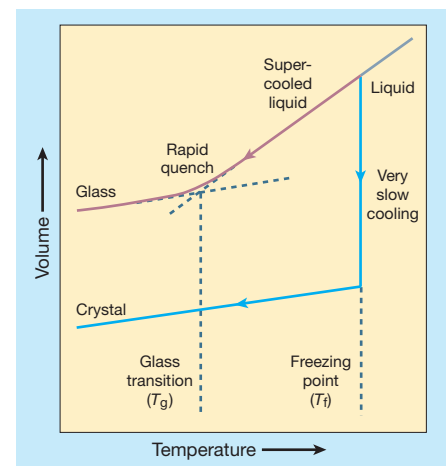


Figure 1 Two cooling paths by which a liquid may solidify. A very slow cooling rate leads to a discontinuous change in volume to a crystal state (purple curve). A rapid quench leads to a continuous change in volume to a glass (blue curve). For a model system of hard spheres, Santen and Krauth¹ have shown that the glass transition does not involve an underlying thermodynamic phase transition.

creating an 'entropy crisis.' The crisis is averted in practice by the intervention of the glass transition. In contrast, the kinetic viewpoint explains the observed structural changes as the consequence of a dynamic transition in the relaxation of the supercooled liquid, which is not accompanied by abrupt changes in thermodynamic properties.

To understand the glass transition, investigators (including Santen and Krauth) have studied idealized models in which the atoms, represented by spheres or disks, interact through hard-core repulsions only⁴. Because the interaction potential is either zero or infinite, such systems are athermal — that is, the thermodynamic and kinetic properties are (after a trivial rescaling) independent of temperature. One might well question whether such a simplified model can shed any light on the nature of the glass transition in real materials. Interestingly, identical hard spheres are known to undergo a transition from a disordered, liquid-like state to an ordered, crystal-like state at a high enough density⁵. Thus, increasing the density in hard-sphere systems plays the same role as decreasing the temperature in thermal systems. Indeed, by increasing the density of hard-sphere systems quickly enough, crystallization can be avoided (albeit only for short periods of time⁶), in analogy with the 'supercooled' branch of the thermal system in Fig. 1.

Computer simulations have led to two schools of thought concerning the glass transition in systems of equal-sized hard spheres. One proposes that the transition is thermodynamic^{4,7} because there are indications of a discontinuity in the slope of the equation of state along the supercooled branch. The

Glass transition

Hard knock for thermodynamics

Salvatore Torquato

It was once thought that relatively few materials could be prepared as amorphous (or disordered) solids. It is now widely believed that the amorphous state is a universal property of condensed matter, whether ceramic, polymeric or metallic¹. The amorphous solid known as a 'glass' can be achieved by quenching (cooling) a liquid sufficiently rapidly to below its glass transition temperature, T_g , to avoid crystallization (Fig. 1). Roughly speaking, a glass is a material that is out of equilibrium, having the disordered molecular structure of a liquid and the rigidity of a solid. But the underlying physics of the glass transition remains one of the most fascinating open questions in materials science and condensed-matter physics. A hotly debated issue is whether the glass transition involves an underlying

thermodynamic (static) or kinetic (dynamic) phase transition. On page 550 of this issue, Santen and Krauth² provide further evidence that the glass transition is not thermodynamic in origin.

A thermodynamic phase transition must involve abrupt changes in certain thermodynamic properties, such as volume. According to the thermodynamic viewpoint, the experimentally observable glass transition is a kinetically controlled manifestation of an underlying thermodynamic transition. This explanation originates with the famous work of Kauzmann³. He observed that the entropy (structural disorder) of a supercooled liquid would fall below that of the corresponding crystal structure — which is expected to have the lowest entropy — if cooled below what is referred to as the Kauzmann temperature,



100 YEARS AGO

A remarkable paper, by Mr. Nikola Tesla, appears in the June number of the *Century Magazine*. The subject is "The Problem of Increasing Human Energy, with Special Reference to the Harnessing of the Sun's Energy"; and though metaphysical and sociological questions receive a large share of attention, the article contains an account of some very interesting electrical experiments... Electrical discharges capable of making atmospheric nitrogen combine with oxygen have recently been produced. Experiments made since 1891 showed that the chemical activity of the electrical discharge was very considerably increased by using currents of extremely high frequency or rate of vibration... The flame grew larger and larger, and its oxidising action more and more intense. From an insignificant brush discharge a few inches long it developed into a marvellous electrical phenomenon, a roaring blaze, devouring the nitrogen of the atmosphere and measuring sixty or seventy feet across. The flame-like discharge visible is produced by the intense electrical oscillations which pass through the coil and violently agitate the electrified molecules of the air. By this means a strong affinity is created between the two normally indifferent constituents of the atmosphere, and they combine readily, even if no further provision is made for intensifying the chemical action of the discharge.

From *Nature* 31 May 1900.

50 YEARS AGO

The subject of population growth or decline has been in the forefront since the Royal Commission on Population began its investigations and deliberations in 1944. Last year its final report was published, and one of its main conclusions was that the slackening of population growth was due to a change in respect of average family size. The Royal Commission outlined the historical background against which the small family came into being: the decline in the importance of the family as a productive unit, the lengthening period during which children remained dependent upon their parents, the advantages of small family size for those struggling for social security and promotion, the increasing and more widely spread knowledge of the techniques of controlled fertility, the changing status of women and the increased opportunities for the enjoyment of leisure outside the home.

From *Nature* 3 June 1950.

other claims that there is no evidence for a thermodynamic glass transition because no such discontinuities are perceivable and the system always crystallizes along the super-cooled branch⁶, implying that (if it exists) the glass transition can only be a dynamical or kinetic phenomenon. Recent experimental work on colloidal hard spheres⁸ supports the second claim.

Adding to the view that glass formation is not a thermodynamic phase transition is the Monte Carlo simulation of Santen and Krauth². They cleverly chose to study two-dimensional systems of hard disks in which each particle has a different size to ensure that the systems do not crystallize. By eliminating any crystallization effects, they can focus exclusively on the nature of the disordered branch. Along this branch, they define the glass transition density to be the point at which the diffusion coefficient (a measure of the mobility of the atoms) becomes vanishingly small — below this density the atoms are mobile (equilibrium liquid), and above this point they are immobile (glass).

An underlying thermodynamic phase transition would be reflected in a discontinuous change in certain thermodynamic properties in crossing this density. In doing these calculations, one must ensure that all of the phase space is sampled without bias (that is, sampling is ergodic). But standard Monte Carlo techniques are known to be non-ergodic when the dynamics slow down near phase transitions. To circumvent this problem, Santen and Krauth use a 'cluster' Monte Carlo algorithm — a method originally introduced to study so-called spin systems

near their critical points^{9,10}. In particular, by a non-local swapping of large clusters of disks, they avoid the problems that conventional Monte Carlo methods have near critical points and find no evidence for a thermodynamic glass transition. Although the specific case studied here does not settle the issue once and for all (for example, other realistic models for glass formation might behave differently), it is another piece of evidence that adherents of the thermodynamic phase-transition theory will find difficult to discount.

There are several ways in which this work can be taken forward. For example, it could be extended to less idealized systems, such as those in which crystallization is not explicitly suppressed and systems with more realistic interaction potentials (allowing for temperature dependence). The possibility of combining the non-local character of the cluster Monte Carlo method with molecular dynamics simulations in order to study the true dynamics of glasses is tantalizing. ■

Salvatore Torquato is in the Department of Chemistry and the Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA.

e-mail: torquato@princeton.edu

1. Debenedetti, P. G. *Metastable Liquids* (Princeton University Press, 1996).
2. Santen, L. & Krauth, W. *Nature* **405**, 550–551 (2000).
3. Kauzmann, W. *Chem. Rev.* **43**, 219–256 (1948).
4. Woodcock, L. V. *J. Chem. Soc. Faraday Trans.* **72**, 1667–1672 (1976).
5. Hoover, W. G. & Ree, F. H. *J. Chem. Phys.* **49**, 3609–3617 (1968).
6. Rintoul, M. D. & Torquato, S. *Phys. Rev. Lett.* **77**, 4198–4201 (1996).
7. Speedy, R. J. *J. Chem. Phys.* **100**, 6684–6691 (1994).
8. Kegel, W. K. *Langmuir* **16**, 939–941 (2000).
9. Swendsen, R. H. & Wang, J.-S. *Phys. Rev. Lett.* **58**, 86–88 (1987).
10. Wolff, U. *Phys. Rev. Lett.* **62**, 361–364 (1989).

Cognitive neuroscience

Learning how the brain learns

Eilon Vaadia

Methods for monitoring brain activity in behaving animals and humans are being developed with increasing pace. But we are still far from understanding the processes in the brain that give rise to even simple types of behaviour. On page 567 of this issue¹, Laubach and colleagues address the problem by recording the activity of many nerve cells as rats learn a simple task, and combine this approach with sophisticated analytical algorithms.

It is widely accepted that large areas of cortex are involved in any behavioural process, and that these areas contain many modules, each consisting of groups of cells that process specific information. It is often assumed that, once the brain matures, each module and each cell fulfils one specific function. But accumulating evidence indicates

that this may not be so. Instead it is likely that each cell participates in several different processes. The brain is also constantly changing, and each cell's effects may be rapidly modified². So it is essential to study a large number of neurons simultaneously to understand how cells communicate and how neuronal interactions are modified in relation to learning and behaviour.

These ideas have been evolving for many years^{3–5}, but concrete facts have been hard to come by. For a start, it is difficult to record the activity of a large number of neurons while still identifying and isolating each cell. Other problems stem from the limited capacity of computer hardware and a lack of suitable data-analysis algorithms. The study by Laubach *et al.*¹ reflects the intense work being done worldwide to tackle these issues.

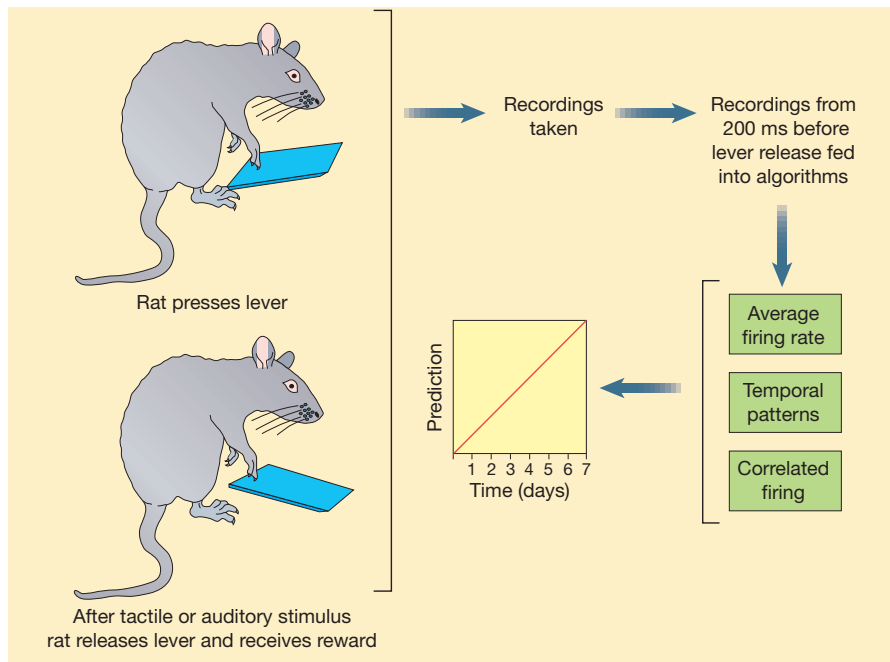


Figure 1 Brain and behaviour. Laubach *et al.*¹ have attempted to understand how the rat brain functions during a simple learning task (left). The information (shown in green) extracted by the artificial neural networks allowed Laubach *et al.* to predict, with increasing certainty over time, whether or not the rats would perform their simple task correctly — see graph (which does not show actual data).

The authors test the hypothesis that information is represented by different groups of neurons, which change their activities and their interactions on both short and longer timescales. Learning-dependent changes in synaptic efficacy that develop on a short timescale (from a fraction of a second to many minutes)⁶ have been studied intensively. Mitz *et al.*⁷ elegantly demonstrated learning-dependent modification of single-cell activity, as a higher animal (a monkey) learnt to associate an arbitrary stimulus with an appropriate movement. Laubach *et al.*, by contrast, follow both the firing rates of single neurons and the interactions within a group of cells during a simple, but longer, learning process.

The trick was to implant microwires in the cortex, rather than inserting micro-electrodes into one site in each 1–3-hour recording session and removing them afterwards. Laubach *et al.* could then follow the development of activity in groups of cells for several days, as a naive rat learnt to press a lever and wait until it was instructed to release the lever. This seems an important achievement — if we can understand the information carried by the neuronal activity.

What is the meaning of ‘understanding’ in this context? The approach taken by Laubach and colleagues, as in several previous studies^{8,9}, was to test whether it is possible for us to predict elements of the rats’ behaviour by extracting information from the spike activity — that is, the patterns of electrical output — of the sampled cells. Their data analysis aimed to answer one

question: how well can we predict the outcome of a single trial (whether or not the rat released the lever at the correct time) from a certain feature of neuronal activity? The authors analysed three features of the spike activity occurring during a period of 200 milliseconds before the start of lever release. These features were the firing rates of all neurons, averaged across the 200-millisecond period; the precise times at which single neurons fired (the temporal firing pattern); and the correlated neuronal activity (Fig. 1).

Analysing the first two of these features is quite simple. The results show — as predicted by intuition and previous studies (for example, ref. 10) — that averaging the firing rate across the 200-millisecond period decreases the amount of available information in the spike trains, as compared with the information carried by the temporal patterns over brief, 10-millisecond intervals. This is not too surprising: many cognitive processes can be accomplished within a short period lasting a few tens of milliseconds.

Studying correlated activity is much more complex, for there are many problems when attempting to extract information from groups of cells — about 20–30 cells in this study. The analysis by Laubach *et al.* is based on complicated procedures, which represent an attempt to interpret data that are complex because of their high dimensionality. To reduce the data representation to manageable levels, the authors first represent the spike trains by the ‘principal

components’ procedure. They then apply ‘independent-component analysis’, a method usually used to separate different sources of information from a mixed signal (for example, to extract speech from background noise). They use this technique, and other algorithms, to estimate how much each cell contributes to each independent component, and to what extent the accurate correlated activation of different cells contributes to the available information about animal behaviour.

The results show that information may be encoded in groups of cells, by both rate and temporal codes, as suggested previously^{11,12}. Laubach *et al.* make significant progress in deciphering these codes on a trial-by-trial basis. This is important, particularly given the idea that neuronal interactions may change rapidly. The authors also show that both firing rates and temporal patterns change during learning. This suggests that neuronal groups are not rigid entities. On a short timescale, groups of neurons are rapidly activated at the same time as the animal prepares its motor response (see Fig. 3 of the paper on page 569). Cells can contribute to different ‘independent components’ by firing in concert at different times. On a longer timescale, new groups organize and tune themselves to represent an arbitrary ‘mapping’ (selected by the researcher) between a sensory cue and a movement. This mapping seems to be highly distributed: in every animal, a small group of 20–30 neurons, spread over quite a large volume of cortex, can be used to predict the behavioural outcome of single trials.

This work provides hope that such approaches will lead to new insights into the mechanisms by which the brain controls behaviour. But there are still issues to resolve. For example, our ability to monitor the same cells for a long period will be improved in the future. Relating the observed modification of neuronal activity to learning per se is difficult, despite the controls used. Interpreting the results of the independent-component analysis is also problematic. Laubach *et al.* used a series of transformations that will leave some readers, including experts, with non-intuitive ideas about the neuronal activity. And, although the methods reveal relevant information about behaviour, it is not possible to test whether the independent components represent independent physiological sources of information.

We are making only the first, tentative steps in a long ‘journey to the brain’. Our progress so far might be compared to that of the Wright brothers, who flew the first aeroplane — if their goal were to reach the moon. Time will tell whether or not research such as that by Laubach and colleagues takes us along the right track. But we should continue to pursue even more ambitious attempts to understand how neuronal networks embody

a computation that can bring about meaningful behaviour. ■

Eilon Vaadia is in the Department of Physiology, Hadassah Medical School and the Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem, Israel.
e-mail: eilon@hbf.huji.ac.il

1. Laubach, M., Wessberg, J. & Nicolelis, M. A. L. *Nature* **405**, 567–571 (2000).
2. Vaadia, E. *et al.* *Nature* **373**, 515–518 (1995).
3. Hebb, D. *The Organization of Behaviour. A Neurophysiological Theory* (Wiley, New York, 1949).
4. von der Malsburg, C. Internal Report 81-2. Max-Planck-Institute for Biophysical Chemistry, Göttingen. Reprinted in

Models of Neural Networks Vol. II (eds Domany, E., van Hemmen, J. L. & Schulten, K.) 95–119 (Springer, Berlin, 1981).

5. Abeles, M. *Local Cortical Circuits: An Electrophysiological Study* (Springer, Berlin, 1982).
6. Byrne, J. H. & Kandel, E. R. *J. Neurosci.* **16**, 425–435 (1996).
7. Mitz, A. R., Godschalk, M. & Wise, S. P. *J. Neurosci.* **11**, 1855–1872 (1991).
8. Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. *Science* **233**, 1416–1419 (1986).
9. Schwartz, A. B. & Moran, D. W. *J. Neurophysiol.* **82**, 2705–2718 (1999).
10. Richmond, B. J. & Optican, L. M. *J. Neurophysiol.* **57**, 147–161 (1987).
11. Gray, C. M., König, P., Engel, A. K. & Singer, W. *Nature* **338**, 334–337 (1989).
12. Maynard, E. M. *et al.* *J. Neurosci.* **19**, 8083–8093 (1999).

Quantum decay

A watched pot boils quicker

Peter W. Milonni

Repeated observations of an unstable quantum state — such as a radioactive atom — can make it live forever, or at least much longer than its natural lifetime. This ‘watched pot’ or ‘quantum Zeno effect’ has been studied intensively over the past decade. Writing on page 546 of this issue, however, Kofman and Kurizki¹ suggest that a ‘quantum anti-Zeno effect’ — the shortening of a lifetime due to repeated observations — may be more common.

An unstable state evolves in time into a linear superposition of states. For instance, an excited state of an atom in a vacuum evolves into a superposition of itself and the (stable) states in which the atom is unexcited and has released a photon into the surrounding space. A measurement to determine whether the initial state survives can be formally described as a projection of the superposition back onto the initial state. The quantum Zeno effect can occur when measurements are repeated so rapidly that the time between them is much shorter than the natural lifetime, or ‘coherence time’, of the state. Coherence times are typically so short that this condition is not satisfied.

Ten years ago² an experiment to test the predicted Zeno effect used a radio-frequency field to excite atoms from a state 1 to a state 2, plus a sequence of laser pulses inducing transitions from state 1 to a higher excited state 3. The detection of a spontaneously emitted photon by the $3 \rightarrow 1$ (fast) transition amounts to a measurement of the survival probability of state 1, because spontaneous emission from state 3 can occur only after the atom has already been in state 1, where it can absorb a laser photon. It was found that radio-frequency transitions from state 1 to 2 were diminished by frequent measurements of the survival of state 1. Whether this confirms the quantum Zeno effect is arguable, or perhaps semantic, in that the enhanced

survival of state 1 can be understood as an interference effect — rather than projecting the atom into state 1, the laser pulses (‘measurements’) produce superpositions of states^{3,4}.

Another example where the survival of a quantum state is affected by sequential measurements occurs in the propagation of a photon through N polarization rotators, each of which rotates the initial (‘horizontal’) polarization by the angle $\pi/2N$. Left to themselves, the rotators in series rotate the polarization by $N(\pi/2N) = \pi/2$, that is, from horizontal to vertical. Suppose now that a horizontal polarizer is placed after each rotator. The probability of the photon being transmitted by each of these polarizers is $p = \cos^2(\pi/2N)$, and the probability

of transmission through all of them is $P = p^N \approx 1 - \pi^2/4N$. The probability of absorption is $1 - P \approx \pi^2/4N$. Many (large N) sequential polarization measurements therefore result in a high survival probability of the initial state. Again it is arguable whether this effect embodies the quantum Zeno effect as originally proposed, but it has been used to demonstrate ‘quantum-interrogation’ measurements^{5,6}.

The original prediction⁷ of a quantum Zeno effect referred to a spontaneous decay of an unstable particle or state. It is this general situation, where decay is a consequence of a ‘reservoir’ of possible states to which transitions can occur, that has been revisited by Kofman and Kurizki¹. They focus attention on the dependence of the decay rate on the energy spectrum of the reservoir states and on the energy spread of the unstable state.

It is clear that the decay rate must depend on the spectrum of reservoir states, simply because the decay is due to transitions to these states. Measurements interrupt and randomize the oscillations of the system and, according to the energy–time uncertainty relation, cause an energy spread $\sim h\nu$, where h is Planck’s constant and ν is the frequency of the sequential measurements. This energy spread determines the range of accessible reservoir states and must also affect the decay rate.

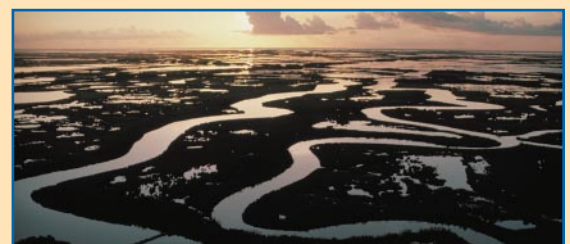
The main conclusions of Kofman and Kurizki follow from an analysis of the dependence of the decay rate on the energy spread and the reservoir spectrum. The quantum Zeno effect should occur when the energy spread due to repeated observations is large compared with both the width of the reservoir spectrum and the separation in

Geomorphology

Case of the bends

Paul Hudson and Richard Kesel have returned to the Mississippi River of Mark Twain to investigate the rates of movement (migration) of river meanders. The inspiration for their study was not literary, however, but old surveys of the 1,700-km lower stretch of the river from Cairo, Illinois, to the Gulf of Mexico. The surveys concerned were carried out between 1877 and 1924 — a time when the river was still largely unconstrained by human influence. The photograph here is a recent one of the river in the delta region.

Data from smaller river



systems and from modelling point to a specific relationship between meander migration rate and the ratio between meander-bend radius and channel width. But as they describe in *Geology* (**28**, 531–534; 2000), Hudson and Kesel find that the lower Mississippi does not follow this relationship. The reason,

they suggest, lies in the heterogeneity of the deposits through which the river flows. In particular, clay plugs limited the rate of meander migration — the incidence of these plugs is higher in the northern part of their study area (average meander migration 45.2 m yr^{-1}) than in the delta (59.1 m yr^{-1}). **Tim Lincoln**

NATHAN BENN/CORBIS

energy between the unstable state and the centre of gravity of the reservoir distribution. In the seemingly more general situation where this energy spread is small compared with the separation between the unstable state and the nearest maximum in the reservoir spectrum, however, the unstable state should decay faster as ν increases. In this case the quantum anti-Zeno effect should be observed because, as ν and therefore the energy spread of the unstable state increases, so does the number of accessible reservoir states into which transitions can occur.

Judging by typical decay processes and the nature of the reservoirs used to model them, one surmises that the anti-Zeno effect is more common than the Zeno effect^{1,8,9}. As yet there are no known experiments corroborating this idea, but it has already been suggested⁸ that the anti-Zeno effect is cause for concern in connection with error-correction

schemes for quantum computing. Moreover, Fearn and Lamb⁴ reported "quite the reverse of the Zeno effect" in a computational study involving barrier penetration. So watch that pot closely. ■

Peter W. Milonni is in the Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

e-mail: pwm@lanl.gov

1. Kofman, A. G. & Kurizki, G. *Nature* **405**, 546–550 (2000).
2. Itano, W. M., Heinzen, D. J., Bollinger, J. J. & Wineland, D. J. *Phys. Rev. A* **41**, 2295–2300 (1990).
3. Ballentine, L. E. *Phys. Rev. A* **43**, 5165–5167 (1991).
4. Fearn, H. & Lamb, W. E. Jr *Phys. Rev. A* **46**, 1199–1205 (1992).
5. Kwiat, P., Weinfurter, H., Herzog, T., Zeilinger, A. & Kasevich, M. in *Fundamental Problems in Quantum Theory* (eds Greenberger, D. M. & Zeilinger, A.) 383–393 (New York Acad. Sci., 1995).
6. Kwiat, P. G. *et al.* *Phys. Rev. Lett.* **83**, 4725–4728 (1999).
7. Misra, B. & Sudarshan, E. C. G. *J. Math. Phys.* **18**, 756–763 (1977).
8. Lewenstein, M. & Rzaewski, K. *Phys. Rev. A* **61**, 022105–1–022105–5 (2000).
9. Kaulakys, B. & Gontis, V. *Phys. Rev. A* **61**, 1131–1137 (1997).

Immunology

The footprint of a killer

Klas Kärre and Gunter Schneider

Class I molecules of the major histocompatibility complex (MHC) are normally expressed on the surface of most cells in the body. They serve two purposes in the immune system. T cells use MHC molecules to identify and kill cells that are infected by foreign material, such as viruses. Natural killer (NK) cells, meanwhile, use them to identify and spare the lives of cells that are normal and healthy¹. On page 537 of this issue², Boyington and colleagues reveal the structural nature of a life-saving liaison between a human MHC molecule and its receptor on an NK cell.

MHC class I molecules bind intracellular peptides and present them to the immune system for scrutiny. If the MHC molecules present foreign peptides — for example, those from a degraded virus protein — T cells are activated and will eventually destroy the infected cells. This limits replication and spread of the infection. NK cells also rely on MHC molecules for decision-making, but in the opposite way. On recognizing self-MHC molecules on target cells (Fig. 1a, overleaf), the MHC receptors on NK cells generate inhibitory signals³. These cancel an activating signal initiated previously by other receptors on NK cells that have recognized ubiquitously expressed ligands on the target cells. If self-MHC molecules are not adequately displayed — as occurs in some cancerous or transplanted cells, for example — the NK receptors cannot deliver the inhibitory signal and will go on to kill the target cell.

Human NK cells can express MHC class I receptors of either the immunoglobulin family (the receptors of which are called killer-cell immunoglobulin-like inhibitory receptors, or KIRs) or the C-type-lectin family. It is the structure of an immunoglobulin-type receptor, KIR2DL2, in complex with its MHC class I ligand, HLA-Cw3, that is described by Boyington *et al.*²

The structure of the extracellular parts of a KIR consists of two globular immunoglobulin-like domains linked by a hinge region⁴. In the MHC–receptor complex described by Boyington *et al.*, the receptor binds with both of its globular domains in a 1:1 stoichiometry to the MHC molecule, across the MHC's 'business end', consisting of the groove-containing peptide. Six loops located close to the hinge region contact the MHC molecule.

Mouse NK cells do not express KIRs. Instead they have inhibitory MHC receptors from the Ly49 subfamily of C-type lectins. At the end of last year, the structure of such a receptor, Ly49A, in complex with its MHC class I ligand, H-2D^d, was described⁵. The interactions between receptor and ligand in this structure differ completely from those described by Boyington *et al.* (Fig. 1b, d), for the mouse receptor binds at one side of the peptide-binding groove. At first sight, however, the major 'footprint' of the human NK-cell receptor on its MHC ligand seems remarkably similar to that of the T-cell antigen receptor on an MHC ligand⁶ (Fig. 1c). Both receptors span the two α -helices of the

MHC molecule and the peptide, covering a surface area of about 1,500–1,700 Å. But a closer examination reveals that the binding site of the NK-cell receptor is shifted towards the carboxy-terminal part of the peptide. The T-cell receptor, by contrast, binds more centrally. Another difference is that the KIR footprint is dominated by charged and hydrophilic interactions, whereas the T-cell receptor relies mainly on hydrophobic, and less on electrostatic, contacts.

The degree of overlap between the two footprints indicates that the T-cell antigen receptor and the KIR probably cannot bind to the MHC molecule at the same time. This is an important issue, as T cells may, under certain circumstances, express both an activating MHC receptor and an inhibitory KIR. Might one MHC molecule be used to deliver both activating and inhibitory signals to the same T cell? This appears impossible for the human KIR studied by Boyington *et al.* (Fig. 1b, c), but might arise for mouse NK cells bearing the Ly49A receptor (Fig. 1b, d).

The role of the MHC-presented peptide in recognition by the NK-cell receptor is much debated. For some receptors, such as mouse Ly49A, there is no influence of the peptide sequence. But the opposite appears to be true⁷ for the group of receptors that includes that studied by Boyington *et al.* This difference is beautifully explained by the three-dimensional structures of the two receptor–ligand complexes^{2,5} (Fig. 1b, d). The binding site of the murine receptor is distant from the peptide. But there are direct interactions between the peptide and the human receptor. Two positions in the non-amer peptide participate in these interactions in the structure — a finding that agrees well with biochemical binding studies using a series of peptide variants².

MHC molecules are polygenic (each individual has several genetic loci that encode MHC molecules) and polymorphic (each locus expresses but one of many possible alleles). Why does one group of KIRs, such as that to which KIR2DL2 belongs, bind to MHC molecules encoded by just one particular locus, termed HLA-C for this group? And why do different KIRs within this group bind to mutually exclusive subgroups of allelic products from this locus? This is usually referred to as the allospecificity of the receptor, believed to be important for the role of NK cells in reactions between donor and host in bone-marrow transplantation⁸. The allospecificity of KIRs has been analysed in several studies⁹, the results of which are explained and extended by Boyington *et al.*'s complex. Of 16 amino acids involved in contacts with the MHC molecule, 14 are conserved between the KIR studied by Boyington *et al.* and another KIR, both of which react with HLA-C ligands but with different allospecificity. Boyington *et al.* found that one of the two amino acids

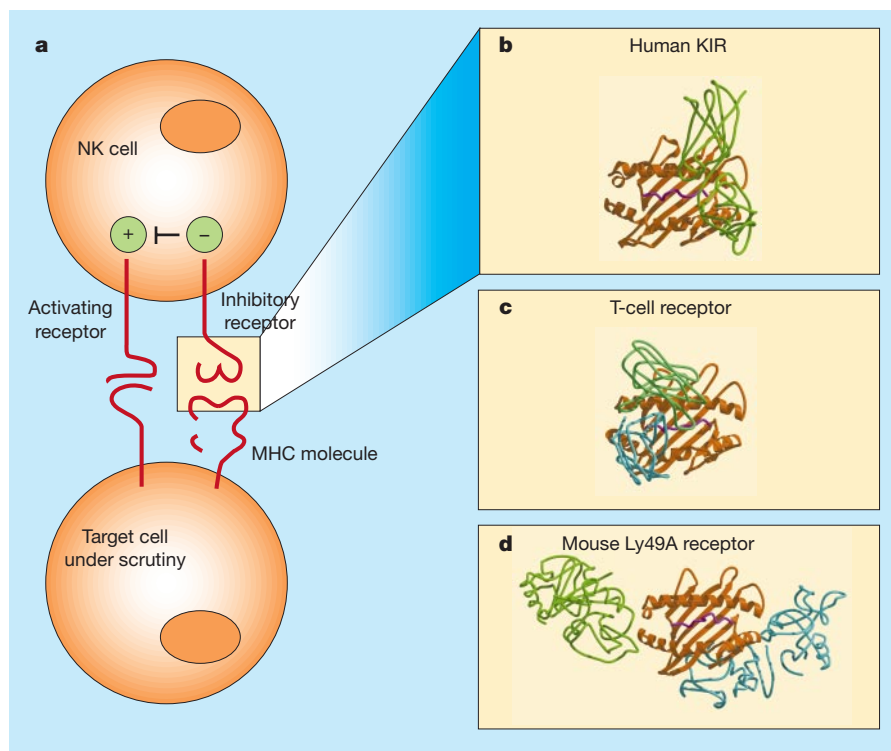


Figure 1 Life-saving liaisons. **a**, Natural killer (NK) cells can distinguish normal, healthy cells from aberrant cells in the body. Activating receptors on the NK cell recognize ligands on most normal cells. Signals from these receptors stimulate the NK cell to kill the target cell. But killing is stopped if killer-cell inhibitory receptors recognize adequate levels on the target cell of major histocompatibility complex (MHC) class I molecules. **b**, The 'footprint' (contact area) made by the human NK receptor KIR2DL2 on the antigen-binding groove of its MHC ligand, HLA-Cw3 (ref. 2), viewed from above. **c**, **d**, The footprints of a T-cell antigen receptor⁶ and of the murine NK receptor Ly49A (ref. 5) are shown for comparison. MHC is shown in red, MHC-bound peptide in purple, and the NK or T-cell receptors in green (and cyan). Note a second contact site for Ly49A beneath the antigen-binding groove. (**b**–**d**, Reproduced with permission from ref. 11.)

that differ between these KIRs is a partner in a critical hydrogen bond with an amino acid that differs between the relevant HLA-C subgroups.

The mouse and human receptor–ligand crystal structures indicate, unexpectedly, another binding site of the receptor to the MHC molecule, not identical in the two complexes. These sites may be involved in ligand-induced receptor aggregation² or receptor interactions with ligands in the membrane of the NK cell itself³. But the biological significance of these interactions is yet to be established.

As ever, questions remain. Why have humans and mice developed different types of NK receptor, with analogous function but from completely different structural families? Might it be because the receptors need to recognize distinct parts of the MHC molecules (Fig. 1b, d)? And if so, why? Structures of other receptor–ligand pairs will be required before we can answer these questions. Genomics studies indicate that these evolutionarily recent KIR proteins show considerable diversity, even at residues that do not contact MHC molecules¹⁰ (P. Parham, personal communication). This diversity is evolving rapidly, even by comparison to the

MHC genes¹⁰. Might peptides or other KIR ligands from infectious agents represent the evolutionary driving force here? Finally, NK cells sometimes carry variants of KIR proteins that activate killing when they recognize MHC ligands. The function of the activating receptors is unknown — will their footprints on the MHC molecule provide the answer? ■

Klas Kärre is at the Microbiology and Tumor Biology Center, Karolinska Institute, Box 280, 17177 Stockholm, Sweden.

e-mail: klas.karre@mtc.ki.se

Gunter Schneider is in the Department of Medical Biochemistry and Biophysics, Karolinska Institute, Box 280, 17177 Stockholm, Sweden.

1. Kärre, K., Ljunggren, H. G., Piontek, G. & Kiessling, R. *Nature* **319**, 675–678 (1986).
2. Boyington, J. C., Motyka, S. A., Schuck, P., Brooks, A. G. & Sun, P. D. *Nature* **405**, 537–543 (2000).
3. Moretta, A. et al. *Annu. Rev. Immunol.* **14**, 619–648 (1996).
4. Fan, Q. R. et al. *Nature* **389**, 96–100 (1997).
5. Tormo, J., Natarajan, K., Margulies, D. H. & Mariuzza, R. A. *Nature* **402**, 623–631 (2000).
6. Garboczi, D. N. et al. *Nature* **384**, 134–141 (1996).
7. Rajagopalan, S. & Long, E. O. *J. Exp. Med.* **185**, 1523–1528 (1997).
8. Ruggeri, L. et al. *Transplantation* **94**, 333–339 (1999).
9. Mandelboim, O. et al. *J. Exp. Med.* **184**, 913–922 (1996).
10. Wilson, M. J. et al. *Proc. Natl Acad. Sci. USA* **97**, 4778–4783 (2000).
11. <http://red.niaid.nih.gov/mhc-complexes.html>

Daedalus

Air currents

Many reactors circulate their molten sodium coolant by an electromagnetic pump. A current passes through the molten metal at right angles to an applied magnetic field, and the motor effect impels the sodium at right angles to both. Daedalus reckons it would work with insulating fluids too. For a voltage applied across an insulator shifts its electrons transiently, as the insulator becomes polarized. Remove the voltage, and they would move back again. In a steady magnetic field, these two brief opposite 'displacement currents' would exert a push and then a cancelling pull on the dielectric. But reverse the magnetic field in the interval, and the dielectric would feel two pushes in the same direction — a true motor effect.

So, says Daedalus, put an insulator in a.c. electric and magnetic fields at right angles, reversing at the same frequency and the right mutual phase, and it will feel a cumulative force. At first he hoped that the whole thing could be driven as a resonant circuit, with the insulator as the dielectric of its capacitor, and the inductor providing the magnetic field; but the phasing comes out wrong. The two fields will have to be created and phased by special circuitry.

The obvious working fluid for the new pump is air. Fans, propellers, blowers and jet engines are all noisy, complicated devices. A simple silent air-pump with no moving parts would be widely welcomed. For maximum thrust, it should work at the highest feasible frequency — many megahertz if possible. Daedalus's first product will be a little cooling blower for computers and electronic gadgets. But he soon hopes to scale it up. Displacement-current pumping could make all sorts of blowers, compressors, air-conditioners and fans blissfully silent and reliable.

His ultimate goal is a displacement-current aircraft engine. The entire craft would have to be designed around this radical new source of thrust. Its fields would enclose as large a volume of air as possible, so that even a gentle induced flow moved a lot of air. An electromagnetic helicopter with vertical downflow might fill the bill. The strength and direction of thrust could be altered at electronic speed, giving the craft a wonderful agility in the air. And its total silence would be welcomed both by the crew and the long-suffering aerophobic public. **David Jones**

The Further Inventions of Daedalus is published by Oxford University Press.

Pacific leatherback turtles face extinction

Fisheries can help avert the alarming decline in population of these ancient reptiles.

The dwindling numbers of leatherback turtles are signalling a threat to biodiversity in the oceans. A mathematical model based on our assessment of a once-large leatherback population predicts that unsustainable adult mortality, apparently due to human fishing activity, will soon drive this population to extinction.

In 1982 there were 115,000 adult female leatherbacks in the world, but in 1996 there were only 34,500 (ref. 1). However, many accounts of decline were based on anecdotal information and indirect measurement. Individuals could not be reliably identified over a period of many years, and the reproductive effort of individual turtles was seldom determined during a given year. Therefore, although leatherbacks had disappeared from India before 1930², declined to near zero in Sri Lanka by 1994³, and fallen from thousands to two in Malaysia by 1994⁴, we did not have enough data to predict the fate of other colonies.

Since 1988 we have studied leatherbacks (Fig. 1) at Playa Grande, Costa Rica, the fourth largest nesting colony in the world¹. Since 1993 we have permanently identified female turtles by injecting them with passive integrated transponder tags. We encountered more than 95% of nesting turtles each night⁵. This enabled us to determine how many nests a turtle laid, how many individual turtles nested in a given year, and how many turtles returned to nest in later years.

In 1988–89 (July–June), 1,367 leatherbacks nested on Playa Grande. By 1994–95, 506 turtles nested there⁵; in 1998–99, there were only 117 (Fig. 2a). In 1996–97, 1997–98 and 1998–99, only 26.7, 27.1 and 20.6%, respectively, of turtles were remigrants. Only 11.9% of turtles tagged in 1993–94 and 19.0% tagged in 1994–95 returned to nest in the next 5 years, with the peak return being in the third year. This population was in the midst of a collapse. In contrast, at St Croix in the Caribbean, remigrants averaged 48.5% from 1989 through to 1995 and the population grew⁶.

How could the Playa Grande turtles have vanished? They could have died; they could still be in the ocean and nesting less frequently; or they could have nested elsewhere. The second explanation is possible but unlikely, as we have determined the mean re-nesting interval to be 3.7 years. Of the 15.4% of leatherbacks that returned to nest in subsequent seasons, 0.5% returned after 1 year, 20.3% returned after 2 years, 46.5% returned after 3 years, 15.5% returned after 4 years, and 17.2% returned



Figure 1 Leatherback turtle (*Dermochelys coriacea*): once abundant in the Pacific, populations have plummeted as a result of capture by fisheries.

after 5 years. Other studies indicate that over 91% of leatherbacks have remigration intervals of 5 years or less^{7,8}. The third explanation is also unlikely because aerial surveys from Mexico to South America have not revealed any other major nesting beaches (R.D.R. and J.R.S., and P. Dutton, unpublished results) and no new major nesting beaches have been reported since 1982¹.

Mortality therefore seems to be the best explanation for the population decline.

From 1993–94 to 1998–99, the annual mortality for Playa Grande leatherbacks that nested in 1993–94 was 34.6%, and for those that nested in 1994–95 it was 34.0%. This is much higher than previously estimated¹ and could be responsible for the rapid decline of this population.

Our model predicts that, without protective measures, the population will fall to less than 50 nesting females by 2003–04 (see Box and Fig. 2b). If beach and hatchery protection is continued, the fall to under 50 animals could be postponed by 5 years. Recovery of this population cannot be achieved by increasing hatchling production alone — even with total protection of beaches, any population suffering these rates of adult mortality cannot survive for more than a few years.

The situation at Playa Grande is reflected at many other Pacific nesting beaches. The large Mexican nesting colony declined exponentially from 70,000 in 1982¹ to under 1,000 by 1994⁹ and to fewer than 250 in 1998–99 (S. Eckert, unpublished results). The annual mortality between 1984 and 1996 was 22.7% (ref. 9).

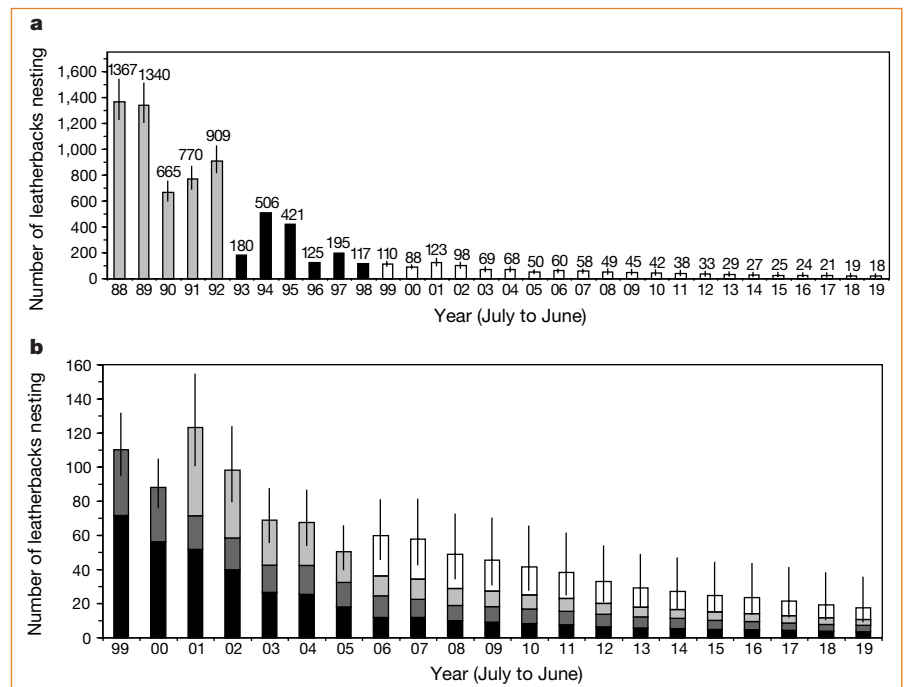


Figure 2 Number of leatherback turtles nesting on Playa Grande, Costa Rica, from 1988–89 to 2019–20. **a**, Stippled bars, numbers based on nest counts and estimated clutch frequency (ECF); black bars, numbers based on individual tagging of leatherbacks with passive integrated transponders. Predictions (white bars) are from the mathematical model (see Box, overleaf). For 1988–92, error bars represent an ECF of 6.0 to 7.5. For 1999–2019, error bars represent effect of maximum and minimum offspring-to-adult ($O:A$) ratio. **b**, Predictions of the number of leatherbacks likely to nest on Playa Grande are based on the mathematical model, the number of leatherbacks nesting in the past, and the effect of protecting the beach. Black portions of bars represent new turtles; dark grey, remigrants; light grey, new leatherbacks resulting from beach protection since 1993–94; and white, new leatherbacks resulting from a hatchery started in 1998–99. Error bars represent effect of maximum and minimum $O:A$ ratio.

A model of the leatherback population on Playa Grande

The model is based on data from previous years and predicts the number of emigrants using the equation

$$R = \sum_{n=1}^{n=5} (r_n \times X_{y-n})$$

where R is the number of turtles returning in year y from previous years, y is the year of prediction, n is the number of years before y , r_n is the decimal fraction of turtles from year $(y-n)$ that return in year y , and X is the number of turtles nesting in a previous year. The model predicts the number of new turtles by comparing the number of recruits nesting in a given year with the number of turtles that nested 6, 7, 8, 9 and 10 years previously, assuming

that new turtles were adult offspring of turtles nesting 6 to 10 years earlier¹¹. We calculated this offspring-to-adult ($O:A$) ratio across the range of years for which we had data, and determined the mean $O:A$ by using the equation

$$O:A = \bar{X}_{m=10}^{m=6} (N_y / O_{y-m})$$

where $\bar{X}$ is the mean ratio for year y , O is the number of turtles nesting in the previous year, m is the age at sexual maturity, N is the number of new turtles, and y is a given year. We computed the overall mean $O:A$ ratio for different ages of maturity in order to apply a single $O:A$ ratio for a predicted year. Mean,

maximum and minimum $O:A$ ratios were 0.093, 0.120 and 0.073. Model predictions for the total number of turtles in 1997–98 and 1998–99 were within 18% of the actual number of turtles counted in those years. The model includes a feedback function to incorporate predicted data in calculations of turtle numbers until the year 2020. We did not have historical data to calculate the $O:A$ ratio for age at sexual maturity of more than 10 years. We assumed that beach protection doubles turtle recruitment and that establishing hatcheries quadruples it.

20036, USA

‡Department of Biology, Indiana-Purdue University, Fort Wayne, Indiana 46805, USA

1. Spotila, J. R. *et al.* *Chelonian Cons. Biol.* **2**, 209–222 (1996).
2. Smith, M. A. *The Fauna of British India, including Ceylon and Burma. Reptilia and Amphibia Vol I. Loricata, Testudines* (Taylor & Francis, London, 1931).
3. Deraniyagala, P. E. P. *The Tetrapod Reptiles of Ceylon Vol I. Testudines and Crocodilians* (Ceylon Government Press, Colombo, 1939).
4. Chan, E. H. & Liew, H. C. *Chelonian Cons. Biol.* **2**, 196–203 (1996).
5. Steyermark, A. C. *et al.* *Chelonian Cons. Biol.* **2**, 173–183 (1996).
6. McDonald, D. L. & Dutton, P. H. *Chelonian Cons. Biol.* **2**, 148–152 (1996).
7. Hughes, G. R. *Chelonian Cons. Biol.* **2**, 153–158 (1996).
8. Boulon, R. H., Dutton, P. H. & McDonald, D. L. *Chelonian Cons. Biol.* **2**, 141–147 (1996).
9. Sarti, L. M., Eckert, S. A., Garcia, N., Barragan, T. & Barragan, A. R. *Mar. Turtle Newslett.* **74**, 2–5 (1996).
10. Eckert, S. A. *Mar. Turtle Newslett.* **78**, 2–7 (1997).
11. Zug, G. R. & Parham, J. F. *Chelonian Cons. Biol.* **2**, 244–249 (1996).
12. Musick, J. A. (ed.) *Life in the Slow Lane: Ecology and Conservation of Long-Lived Marine Animals* (Am. Fisheries Soc., Bethesda, 1999).
13. Pauly, D., Christensen, V., Froese, R. & Palomares, M. L. *Am. Sci.* **88**, 46–51 (2000).

Conservative estimates are that longline and gill-net fisheries killed at least 1,500 female leatherbacks per year in the Pacific during the 1990s^{1,10}. These included Asian trawl, longline and drift-net, Central and South American longline and gill-net, and Hawaiian longline fisheries. With a population of about 6,500 adult females¹, this corresponds to a 23% annual mortality, or 33% if most leatherbacks captured came from the East Pacific population of 4,600 animals¹. Most of the mortality at Playa Grande was probably caused by fisheries. Leatherbacks normally live at least 30 years and reach maturity at 5–14 years¹¹. A long-lived species like this cannot withstand such high rates of anthropogenic mortality^{1,12}.

From our tagging data and from aerial surveys, we calculate that there are now 687 adult females and 518 subadults in the Central American population; we estimate that the Mexican population stands at 1,000 adult and 750 subadults. The East Pacific leatherback population thus contains about 1,690 adult females, down from 4,638 in 1995¹. The total adult and subadult population is about 2,955 females — compared with over 91,000 adults in 1980¹. We conclude that leatherbacks are on the verge of extinction in the Pacific.

If these turtles are to be saved, immediate action is needed to minimize mortality through fishing and to maximize hatchling production. Assuming leatherbacks can withstand a 1% annual mortality inflicted by humans¹, the East Pacific population can tolerate the annual loss of 17 adult females and 13 subadult females per year. In 1995 there were about 1,800 females in the Western Pacific¹, so anthropogenic mortality in that population from all causes should not exceed 18 adult females per year. We believe that fishing practices¹³ in the Pacific must be changed to save marine biodiversity.

James R. Spotila*, Richard D. Reina*, Anthony C. Steyermark*, Pamela T. Plotkin†, Frank V. Paladino‡

*School of Environmental Science, Engineering and

Policy, Drexel University, Philadelphia, Pennsylvania 19104, USA

e-mail: spotiljr@drexel.edu.

†Center for Marine Conservation, Washington DC

Palaeontology

Fossil record of mass moth migration

The fossil record of moths and butterflies is extremely poor in comparison with other winged-insect groups, with only an estimated 600–700 specimens of fossil Lepidoptera being known¹. Here I report the discovery of huge numbers of lepidopteran fossils (about 1,700 specimens) in marine, diatomous sediments of the Fur Formation from the lowermost Tertiary of Denmark (55 million years old). The abundance of the most common species indicates that mass migrations occurred over the Palaeogene North Sea, so the scant fossil record of Lepidoptera reflects poor preservation and not a paucity of lepidopteran species or individuals during the Tertiary.

The material consists of complete individuals, wingless bodies and isolated wings from at least seven species. More than 1,000 specimens belong to a species with a body length of about 14 mm. The males of this species have a bundle of wing-coupling bristles called a 'composite frenulum', revealing them to be members of the Heteroneura, the group that comprises the majority of lepidopterans.

Individuals of this species are often found embedded closely together in the sediment: one slab, with a diameter of about 150 mm, contained a group of 14 specimens. Given that these were deposited over an offshore area of the Palaeogene North Sea, the high abundance and density of individuals indicates that this species undertook mass migrations. The species'

abundance in different horizons of the marine Fur Formation lends further weight to this idea, and shows that their flights were not a singular or local phenomenon. The great abundance of lepidopteran fossils in the Fur Formation also indicates that the rareness of fossil Lepidoptera in terrestrial deposits is mainly due to taphonomic processes^{2,3}, and that, just as today, they were a dominant insect group in the Palaeogene terrestrial ecosystem.

Several extant species of butterflies and moths migrate across the North Sea^{4–6}. The number of Lepidoptera, as well as other insects, caught 30 km offshore over the North Sea increases when winds are calm and temperatures over land are high⁵: the fossil Heteroneura specimens probably embarked on their migratory flights during similar summer conditions.

The insects found in the Fur Formation came from the former southwestern Scandinavian coast, about 50–100 km from the depositional area, where they lived in a paratropical woodland with stagnant and flowing waters⁷. The insect fauna from the main part of the Fur Formation remains constant through a sediment column of about 30 m. There is no evidence for major climatic or environmental changes, and I believe that the flight of the Lepidoptera and other insects was seasonal.

The otherwise common Heteroneura species is missing from the basal Ølst Formation. Instead, its laminated clay and shale deposits are dominated by poorly flying insects such as giant ants⁸, damselflies and crickets⁹. These indicate inshore conditions, whereas the Lepidoptera, and other insects capable of long-distance flight, found in the overlying sediments of the Fur

Formation, were deposited in an offshore area. To my knowledge, this is the first time that a marine transgression can be traced using the analysis of fossil insects. These results agree with investigations of the fish fauna of the Ølst and Fur Formations¹⁰.

Jes Rust

Institut für Zoologie und Anthropologie,
Universität Göttingen, Berliner Strasse 28,
37073 Göttingen, Germany
e-mail: just@gwdg.de

1. Kristensen, N. P. & Skalski, A. W. in *Lepidoptera, Moths and Butterflies* (ed. Kristensen, N. P.) 7–25 (Handb. Zool. 35, de Gruyter, Berlin, 1999).
2. Martínez-Delclòs, X. & Martinell, J. *Kaupia* 2, 133–144 (1993).
3. Rust, J. *Paläontol. Z.* 72, 41–58 (1998).
4. Williams, C. B. *The Migration of Butterflies* (Oliver & Boyd, Edinburgh, 1930).
5. Heydemann, B. *Deutsche Entomol. Z. (NF)* 14, 185–215 (1967).
6. Johnson, C. G. *Migration and Dispersal of Insects by Flight* (Methuen, London, 1969).
7. Larsson, S. G. *Bull. Geol. Soc. Den.* 24, 193–209 (1975).
8. Rust, J. & Andersen, N. M. *Zool. J. Linn. Soc.* 125, 331–348 (1998).
9. Rust, J. *Entomol. Scand.* 30, 35–45 (1999).
10. Bonde, N. *Aarhus Geosci.* 6, 33–48 (1997).

Neurobiology

Presenilin-1 mutations in Alzheimer's disease

Mutations in the gene encoding the protein presenilin-1 are the most common cause of familial Alzheimer's disease¹ and they often produce a different disease course from sporadic Alzheimer's and another familial form associated with mutations in the gene encoding β -amyloid precursor protein². Here we show that a peculiar form of β -amyloid that is devoid of the first ten amino acids accumulates in the brains of patients carrying presenilin-1 mutations, and is more abundant than in subjects affected by the other types of Alzheimer's.

Carriers of presenilin-1 (PS1) mutations develop a form of Alzheimer's disease (AD) with an earlier onset and shorter disease duration than sporadic cases or patients with familial AD associated with a point mutation (V717I, in which valine and isoleucine are swapped at residue 717) in the β -amyloid precursor protein (APP); they also have unusual clinical symptoms and atypical histological features³. Mutant APP and PS1 may interfere with the normal processing of APP, enhancing the formation of β -amyloid peptides ending at residue 42 ($A\beta_{1-42}$) compared with peptides ending at residue 40 ($A\beta_{1-40}$) (refs 1,3–5). The protease that generates $A\beta_{1-40}$ and $A\beta_{1-42}$ from the full-length protein, γ -secretase, is proposed to be PS1 (ref. 6). However, $A\beta$ peptides trimmed from the other end (amino terminal) may also be an important component of the deposits found in the brains of patients with Alzheimer's^{7,8}.

At least two of the pathogenic PS1 mutations we study here, which give rise to a particularly malignant phenotype, are associated with an excess of amino-terminally truncated β -amyloid and a low concentration of the full-length protein, indicating that mutations of the PS1 gene influence both γ - and β -secretase activities⁹.

We have used immunoblotting to detect the various $A\beta$ species present in brain tissue from subjects with sporadic AD ($n=17$), familial AD linked either to mutations in the PS1 gene ($n=11$) or V717I mutation in the APP gene ($n=2$), and healthy controls ($n=3$). In the soluble fraction prepared from all the diseased brains, $A\beta$ electrophoretically resolves into three bands of relative molecular mass 4.5K (B1), 4.2K (B2) and 3.5K (B3) which are not detectable in the controls (Fig. 1a–c)^{3,8}.

MALDI-TOF mass spectrometry (see Supplementary Information) of $A\beta$ peptides isolated from sporadic AD and PS1-AD brains indicated that the three bands contained $A\beta_{1-40/42}$ (B1); pyroglutamyated $A\beta_{py3-42}$ (B2); and $A\beta_{4-42}$ and pyroglutamyated $A\beta_{py11-42}$ (B3) — in agreement with results obtained using specific antibodies^{7,8}.

Quantitative analysis revealed that the B3

band was significantly ($P<0.003$) more prominent in subjects carrying PS1 gene mutations than in those with sporadic AD or in those with a defective APP gene, again indicating that the amino-terminally truncated forms are increased in PS1 mutants (Fig. 1a,d). Both the B2 and B3 forms were more abundant in patients with the PS1 mutations M139I and H163A (42% and 33% increase, respectively) than in those with sporadic AD (Fig. 1a,d). The $A\beta_{1-42}$ peptide was underrepresented, as demonstrated by using different antibodies raised against $A\beta$ (Fig. 1b,c) in the various cortical regions of members from both families (data not shown).

An excess of amino-terminally truncated $A\beta$ peptides over the full-length protein was also evident in plaques deposited in the brains of PS1-AD patients. Immunostaining of brain sections indicated that, on average, eight times more plaques reacted with anti- $A\beta_{42}$ antibody than with antibody against $A\beta_1$ in carriers of the M139I and H163A mutations (Fig. 1e), although the ratio of plaques recognized by the two antibodies in sporadic AD brains was close to unity (data not shown).

Our results suggest that PS1 mutations affect the cleavage of APP at the amino-ter-

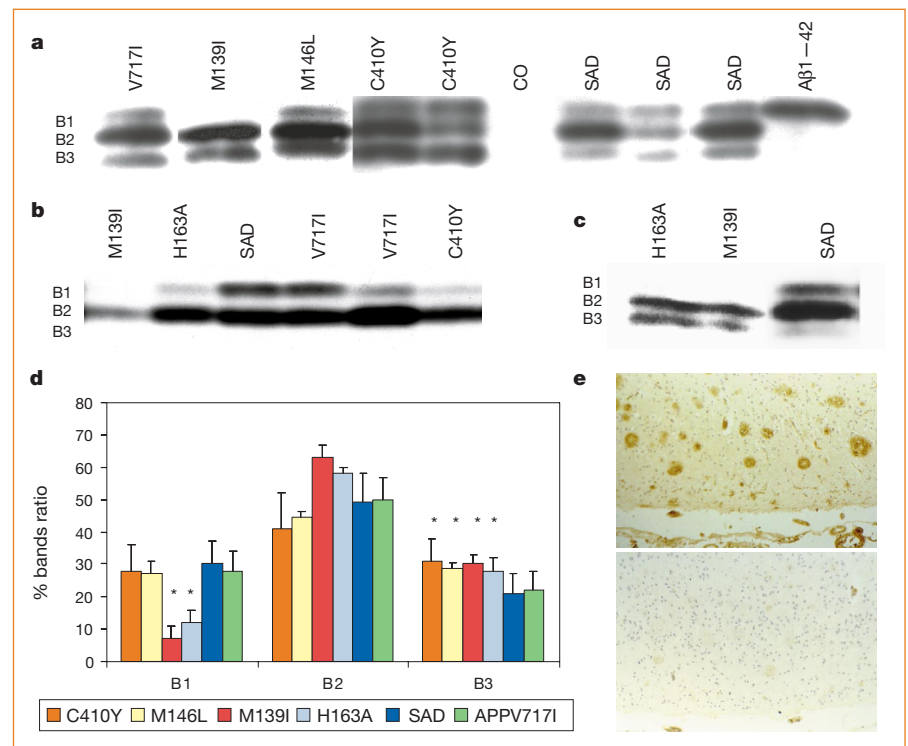


Figure 1 Characterization of peptides derived from β -amyloid precursor protein. **a**, On immunoblots, antibody 4G8 detects three bands in patients with Alzheimer's disease not present in controls (CO). These are B1–3, corresponding to $A\beta_{1-42}$, $A\beta_{3(4)-42}$ and $A\beta_{11-42}$, respectively. In brains of patients with mutant PS1, there is more B3 than in sporadic-Alzheimer's (SAD) patients or in patients with mutant β -amyloid precursor protein. **b,c**, Immunoblotting with antibody 6E10 and another raised against $A\beta_{1-42}$ (which react with residues 6–10 and with alanine 42, respectively) confirms that there is less full-length $A\beta_{1-42}$ produced in PS1 mutants than in sporadic-Alzheimer's patients. **d**, B3 is significantly increased (33–42%) in all PS1 mutants compared with the other types. In the M139I and H163A mutants, B2 is also increased, resulting in a 2.5–4-fold decrease in the full-length forms (asterisk, $P<0.003$). All data are from triplicate experiments. **e**, Eight times more plaques are identified by the antibody against $A\beta$ residue 42 (top) than by the antibody against residue 1 (bottom) in brains with the M139I PS1 mutation. Both antibodies immunostain an equal number of plaques in sporadic Alzheimer's (data not shown).

minal end of the A β sequence. The resulting overexpression of amino-terminally truncated A β peptides indicates that not only is cleavage by γ -secretase affected by PS1 mutation, but that that by β -secretase is as well. PS1 mutation might also affect these secretases indirectly by interfering with the transfer of APP in the cell. The β -secretase BACE¹⁰ preferentially cuts APP at positions 1 or 11 of the A β sequence. Although we do not know how PS1 mutation affects cleavage by β -secretase, on the basis of our results we would expect an increase in BACE cleavage products in PS1-AD subjects. The presence of A β _{py3-42}, A β ₄₋₄₂ and other amino-terminally shortened fragments in these brains indicates that other β -secretases could be involved, or that BACE selectivity might be affected by PS1 mutation. The effect of PS1 mutations on the activity of γ - and β -secretases indicates that PS1 could be involved in the regulation of both secretases, rather than itself being a secretase⁶.

Carriers of PS1 mutations suffer a significantly earlier disease onset (43.5 ± 0.7 years) and shorter dementia (4.5 ± 0.7 years) than patients with mutant forms of APP (61 ± 2.8 and 12.5 ± 2.1 years, respectively) or with sporadic Alzheimer's (75 ± 8.7 and 8.8 ± 4.0 years, respectively). The direct correlation between increased deposits of amino-terminally truncated A β peptides and the clinical course of PS1-AD argues that this is influenced by the presence of these fragments in patients' brains. Similarly truncated A β fragments are overproduced in transfected cells carrying an APP mutant that is associated with an early-onset Alzheimer's phenotype¹¹, and in atypical PS1-AD phenotypes such as the spastic paraparesis variant¹². In prion disease, various species of pathogenic protein are also associated with different phenotypes¹³.

C. Russo[†], G. Schettini[†], T. C. Saido[‡], C. Hulette[§], C. Lippall, L. Lannfelt[¶], B. Ghetti[#], P. Gambetti^{*}, M. Tabaton[✱], J. K. Teller^{††}

^{*}Division of Neuropathology, Institute of Pathology, Case Western Reserve University, Cleveland, Ohio 44106, USA

e-mail: pxg13@po.cwru.edu

[†]National Institute for Cancer Research,

Advanced Biotechnology Center,

Neuroscience Department, Section of Pharmacology, University of Genova, Italy

[‡]Laboratory for Proteolytic Neuroscience, RIKEN Brain Science Institute, Saitama, Japan

[§]Kathleen Price Bryan Brain Bank and Neuropathology Core, Duke University Medical Center, Durham, North Carolina 27710, USA

[¶]Department of Neurology, MCP Hahnemann University, Philadelphia, Pennsylvania 19129, USA

[#]Karolinska Institute, Geriatric Laboratory, Huddinge Hospital, 141 46 Huddinge, Sweden

^{*}Department of Pathology, Indiana University Medical Center, Indianapolis, Indiana 46202, USA

[✱]Institute of Neurology, Genova, Italy

^{††}Present address: Department of Physiology, Brody School of Medicine, East Carolina University, Greenville, North Carolina 27834, USA

1. Selkoe, D. J. *Trends Cell Biol.* **8**, 447–453 (1998).
2. Lippa, C. F. *et al. Neurology* **46**, 406–412 (1996).
3. Teller, J. K. *et al. Nature Med.* **2**, 93–95 (1996).
4. Scheuner, D. *et al. Nature Med.* **2**, 864–870 (1996).
5. Citron, M. *et al. Nature Med.* **3**, 67–72 (1997).
6. Wolfe, M. S. *et al. Nature* **398**, 513–517 (1999).
7. Saido, T. C. *et al. Neuron* **14**, 457–466 (1995).
8. Russo, C. *et al. FEBS Lett.* **409**, 411–416 (1997).
9. Vassar, R. *et al. Science* **286**, 735–741 (1999).
10. Sudoh, S. *et al. J. Neurochem.* **71**, 1535–1543 (1998).
11. Anclio, K. *et al. Proc. Natl Acad. Sci. USA* **96**, 4119–4124 (1999).
12. Crook, R. *et al. Nature Med.* **4**, 452–455 (1998).
13. Bessen, R. A. & Marsh, R. F. *J. Virol.* **68**, 7859–7868 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Astronomy

A massive cool dust torus around η Carinae?

The star η Carinae, the most luminous known in the Milky Way galaxy, is a puzzling object. It is surrounded by a cloud of glowing gas that shows a distinct 'pinched waist' appearance, whose origin has been controversial. Morris *et al.*¹ proposed that a massive but very compact torus of gas encircles the star, and that this caused the bipolar shape by preventing gas ejected by the star from expanding in the equatorial plane. Here we show that the small size of the torus (diameter of 5 arcsec) is insufficient to produce the observed infrared flux that inspired the conjecture. The dust giving rise to the 17- μ m emission must be distributed over an area at least ten times the size of the torus. Such an extended structure of gas, however, would probably be incapable of generating the pinched waist.

As no suitable far-infrared images are available, Morris *et al.* based their hypothesis on features seen at wavelengths around 17 μ m, where only a small fraction of the total flux is due to the cool 110 K dust in their model. The toroidal structure in question, marked in their Fig. 2, is a familiar fea-

ture in previous 2–18- μ m infrared maps of η Carinae and has mid-infrared colour temperatures above 200 K (refs 2–8). A simple brightness–temperature argument shows that it cannot account for much of the cooler radiation seen in the data from the Infrared Space Observatory (ISO). At a wavelength of 50 μ m, for example, the 110 K dust must produce a flux $F_{\nu} \approx 22,000$ Jy in their model (Fig. 1 of ref. 1). In order to have that flux at that wavelength, a 110 K Planck black body would need a projected area of 37 arcsec².

But this is surely an underestimate, because the region must be optically thin at 50 μ m for at least two reasons: first, Morris *et al.* assumed optically thin dust in their analysis of the spectral distribution; second, an opaque torus covering that area would be evident in visual-wavelength images. Thus, one can place a lower limit of roughly 100 arcsec² on the extent of the cool dust component. The projected area of the torus, however, is only 5–10 arcsec². Evidently that structure is far too small to account for the observed far-infrared flux. More complicated arguments involving heating of the grains, 17- μ m intensities in the torus, and so on, lead to the same conclusion.

As Morris *et al.* emphasized, the ISO data seem to indicate at least 5 solar masses of additional ejecta, larger than the amount usually quoted for the bipolar 'Homunculus nebula' of ejecta from η Carinae⁹ (Fig. 1). We suggest that a viable model will have two characteristics. First, a continuous distribution of dust temperatures above 100 K is to be expected. This would fit the observed spectral distribution at least as well as the 110 K and 190 K components assumed by Morris *et al.*, and may reduce the implied mass. Second, and more important, the cool 100–150 K dust is distributed much farther from the star than they proposed. Unshielded dust grains outside the homunculus are expected to have temperatures in that range. Existing far-infrared data provide little or no information on whether the cool dust is equatorial or not. We hope that future instruments will provide images at critical wavelengths beyond 30 μ m.

Kris Davidson, Nathan Smith

Astronomy Department, University of Minnesota, 116 Church Street SE, Minneapolis, Minnesota 55455, USA

1. Morris, P. W. *et al. Nature* **402**, 502–504 (1999).
2. Hyland, A. R. *et al. Astrophys. J.* **233**, 145–153 (1979).
3. Mitchell, R. M., Robinson, G., Hyland, A. R. & Jones, T. J. *Astrophys. J.* **271**, 133–142 (1983).
4. Hackwell, J. A., Gerhiz, R. D. & Grasdaalen, G. L. *Astrophys. J.* **311**, 380–399 (1986).
5. Rigaut, F. & Gehring, G. *Rev. Mex. Astron. Astrofis. Ser. Conf.* **2**, 27–35 (1995).
6. Smith, C. H. *et al. Mon. Not. R. Astron. Soc.* **273**, 354–358 (1995).
7. Smith, N., Gerhiz, R. D. & Krautter, J. *Astron. J.* **116**, 1332–1345 (1998).
8. Polomski, E., Telesco, C., & Piña, R. *ASP Conf. Ser.* **179**, 41–45 (1999).
9. Davidson, K. & Humphreys, R. M. *Annu. Rev. Astron. Astrophys.* **35**, 1–32 (1997).

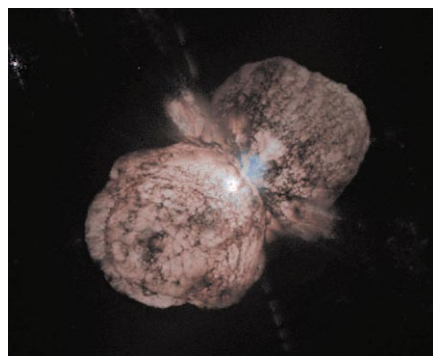


Figure 1 The 'Homunculus' nebula around η Carinae, with axial diameter 17 arcsec, or 0.6 light year.

Inferring the statistical interpretation of quantum mechanics from the classical limit

Kurt Gottfried

Laboratory for Nuclear Studies, Cornell University, Ithaca, New York 14853, USA

It is widely believed that the statistical interpretation of quantum mechanics cannot be inferred from the Schrödinger equation itself, and must be stated as an additional independent axiom. Here I propose that the situation is not so stark. For systems that have both continuous and discrete degrees of freedom (such as coordinates and spin respectively), the statistical interpretation for the discrete variables is implied by requiring that the system's gross motion can be classically described under circumstances specified by the Schrödinger equation. However, this is not a full-fledged derivation of the statistical interpretation because it does not apply to the continuous variables of classical mechanics.

The orthodox interpretation of quantum mechanics asserts that the maximal knowledge available about the state of a system is given by a wavefunction or state vector; that this mathematical object determines the probabilities that specific values of the system's dynamical variables will be revealed by measurement; and that it is not possible, in general, to assign definite values to these variables prior to measurement.

In an important sense, the statistical interpretation remains controversial because the basic dynamical equations of the theory seem not to reveal what the new quantities appearing in them mean in terms of pre-existing concepts, nor do these equations point directly to the orthodox interpretation. As a consequence, the meaning of the wavefunction is traditionally stated as an axiom that stands apart from the Schrödinger equation.

Before the advent of quantum mechanics, physics ultimately proved to be a cumulative pursuit, with new knowledge built on prior concepts. In retrospect, the successful developments of physics followed a clear conceptual path beginning in the *Principia*, but in 1925 this path entered a contested no-man's-land from which it has not yet emerged.

Newton's equations of motion defined the new concepts of mechanics in terms of then accepted conceptions of space and time. Statistical mechanics demonstrated how thermodynamics could be related to an underlying newtonian description. In electrodynamics, the meaning of the new concept of the electromagnetic field is defined in terms of the motion of charged particles by Maxwell's equations, the Lorentz force law and Newton's equations. Admittedly, special relativity introduced a very new conception of space and time, but their meaning was defined by a more penetrating examination of pre-existing concepts using familiar tools: measuring rods and clocks. General relativity, while profoundly new, is firmly tied to pre-existing concepts by its description of phenomena in a local inertial frame. Furthermore, Einstein's equations tell you that—you do not need him whispering in your ear.

Quantum mechanics does not appear to fit this pattern. Whereas Newton (or certainly Hamilton) could have come to understand what is meant by the electromagnetic field had he only been handed Maxwell's equations and the Lorentz force law, could Maxwell have figured out what the wavefunction means had he been handed Schrödinger's equation? It would seem that Maxwell would have needed help from the wonder rabbis of Copenhagen and Göttingen.

My aim here is to explore whether wonder rabbis are indispensable—to ask: To what extent, if any, is the statistical interpretation of quantum mechanics revealed by the theory's formalism and the requirement that quantum mechanics must reduce to classical mechanics under circumstances defined by quantum mechanics itself?

Continuous degrees of freedom

The argument that follows considers systems that have both continuous degrees of freedom, such as coordinates, and discrete variables, such as spin. The Schrödinger equation will be replaced by its classical limit insofar as the continuous variables are concerned, and applied to experiments that discriminate between different values of the discrete variables. It will be shown that this implies the statistical interpretation for the discrete variables, which have no immediate classical counterparts. The argument does not suffice to arrive at the statistical interpretation for degrees of freedom with a continuous spectrum—in particular, the canonical coordinates and momenta, which do have direct classical counterparts. Admittedly, this is a somewhat paradoxical claim.

The argument is presented as a fable featuring a time-warped Maxwell, who is handed the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{q}, t) = \left(\sum_{k=1}^N \frac{1}{2m_k} \left(\frac{\hbar}{i} \frac{\partial}{\partial \mathbf{q}_k} \right)^2 + \sum_{k>j} \frac{e_k e_j}{|\mathbf{q}_k - \mathbf{q}_j|} \right) \Psi(\mathbf{q}, t) \quad (1)$$

and asked to elicit the meaning of Ψ from this equation by exploiting his knowledge of mathematics and classical physics. Maxwell is told that this equation, as it stands and without elaboration, correctly describes all phenomena when examined with a resolution coarser than about 10^{-11} cm, provided the velocities are small compared to c and radiation is negligible. He is also told that $\mathbf{q}_k$, m_k and e_k are the coordinates, masses and charges of the particles that are the constituents of matter, and given the value of $\hbar$.

The stated realm of validity of equation (1) tells Maxwell that this equation must somehow embrace classical mechanics, and that the mysterious constant $\hbar$ must therefore disappear when classical mechanics is reliable, for in that theory the constants e_k and m_k suffice to prescribe all motions. Further, the equation implies that Ψ is complex, so the limit $\hbar \rightarrow 0$ must only allow real quantities to survive.

Facing this puzzle, Maxwell soon proves that the solutions of the equation have an essential singularity at $\hbar = 0$ (ref. 1). This motivates him to make the Ansatz that would later be discovered by Brillouin and Wentzel

$$\Psi(\mathbf{q}, t) = e^{i\Theta(\mathbf{q}, t)/\hbar} \quad (2)$$

where $\mathbf{q} \equiv (\mathbf{q}_1, \dots, \mathbf{q}_N)$ is a point in the configuration space C of the system. This done, he finds that equation (1) produces the following equation for Θ

$$\frac{\partial \Theta}{\partial t} + \sum_k \frac{1}{2m_k} \left(\frac{\partial \Theta}{\partial \mathbf{q}_k} \right)^2 + V(\mathbf{q}) = -\frac{\hbar}{i} \sum_k \frac{1}{2m_k} \frac{\partial^2 \Theta}{\partial \mathbf{q}_k^2} \quad (3)$$

where V is the electrostatic potential in equation (1).

“Aha!” says Maxwell, “this is the Hamilton–Jacobi (HJ) equation if the term proportional to the mysterious constant $\hbar$ is dropped, which also gets rid of the curious i .” (Maxwell does not know that in essence he is trying to reverse Schrödinger’s leap from classical to quantum mechanics².)

So his problem has now become, first, that of learning whether this relation to the HJ equation unlocks the meaning of Ψ ; and second, that of analyzing the departures from the HJ equation implied by the mystery equation. According to equation (3), these departures will be small if Θ satisfies

$$|\partial \Theta / \partial \mathbf{q}_k|^2 \gg \hbar |\partial^2 \Theta / \partial \mathbf{q}_k^2| \quad (4)$$

which is thus the condition under which the new theory can be replaced by the classical HJ equation. However, were Θ to be only replaced by a solution $S(\mathbf{q}, t)$ of the HJ equation, the meaning of Ψ would remain mysterious, for $S(\mathbf{q}, t)$ is real in the classically accessible regions C_{class} of configuration space, and $\exp(iS/\hbar)$ would just be a phase that oscillates ever more violently as $\hbar \rightarrow 0$. However, equation (4) suggests that S is the first term in an expansion of Θ about $\hbar = 0$, so Maxwell tries

$$\Theta = S + (\hbar/i)W + (\hbar/i)^2 X + \dots \quad (5)$$

or

$$\Psi(\mathbf{q}, t) = e^{W(\mathbf{q}, t)} e^{i[S(\mathbf{q}, t)/\hbar + O(\hbar)]}. \quad (6)$$

Maxwell is astonished that the leading correction to the classical approximation gives a factor e^W in Ψ that does not depend on $\hbar$, which prompts him to believe that this factor must be accessible to his understanding. He easily shows that W is real in C_{class} , and with more difficulty that the equation relating W to S can be cast into the suggestive form

$$\frac{\partial w}{\partial t} + \sum_k \frac{\partial}{\partial \mathbf{q}_k} (w \mathbf{q}_k) = 0 \quad (7)$$

where

$$w(\mathbf{q}, t) = e^{2W(\mathbf{q}, t)} = \lim_{\hbar \rightarrow 0} |\Psi(\mathbf{q}, t)|^2 \quad (8)$$

$$\dot{\mathbf{q}}_k(t) = \frac{1}{m_k} \frac{\partial S}{\partial \mathbf{q}_k} \quad (9)$$

the latter being the velocity of particle k when it is moving in accordance with the classical equation of motion and at the position $\mathbf{q}_k(t)$.

Maxwell of course recognizes equation (7) as the continuity equation for a classical fluid of density $w(\mathbf{q}, t)$ and local velocity $\dot{\mathbf{q}}(t)$ flowing in configuration space—not three-dimensional space. Furthermore, as he knows that the classical trajectories are the curves in C_{class} everywhere normal to the surfaces of constant S , he realizes that Ψ is related not to one classical trajectory, but to a whole set of such trajectories, $\{\mathbf{q}(t)\}$.

This then leads Maxwell to conclude that in the classical limit the Schrödinger equation does not describe a single system, but a population of replicas of such a system moving along the set of

trajectories $\{\mathbf{q}(t)\}$. As Ψ is only a function of the degrees of freedom of a single system, and because experiments can be done on individual systems, he visualizes this population as a set of identical specimens which, one at a time, follow one or another of the allowed classical trajectories, and which are populated according to the density $w(\mathbf{q}, t)$. Maxwell also shows that as the masses in Schrödinger’s equation increase, $w(\mathbf{q}, t)$ can be confined to an ever smaller region and for an ever longer interval without violating the condition (4) for the validity of the approximation (6).

For any specific solution Ψ of a specific Schrödinger equation, that is all that Maxwell can say about the trajectories and their population. To have consistently better knowledge of which trajectories are being followed he would have to change the hamiltonian in Schrödinger’s equation by hand so as to confine the trajectories to an appropriately small region, but in so doing he would alter Ψ dramatically.

This fable is not a derivation of the Born interpretation of $|\Psi|^2$ as the probability of measurement outcomes. To emphasize this, I have spoken of ‘populations’, not probabilities. Furthermore, the classical description is only valid when the inequality (4) is satisfied, that is, if the system is sufficiently heavy, the forces sufficiently smooth, and the time interval over which the classical description is to be used is sufficiently short. The Born interpretation of the wave function in configuration space has no such restrictions, quite aside from the profound difference between quantum mechanical probabilities on the one hand, and on the other the probability distribution for a population of replicas obeying the laws of classical mechanics each of which traverses a fully knowable and predictable trajectory.

Indeed, Maxwell soon realizes that his interpretation fails in many situations because of the linearity of the mysterious equation. A superposition of its solutions, $\Psi = \sum_{\alpha} c_{\alpha} \psi_{\alpha}$, when the approximation (6) is made for each term, will result in an $\hbar$ -dependence of $|\Psi|^2$ because of the interference terms

$$\text{Re } \psi_{\alpha}^* \psi_{\beta} \approx \exp[W_{\alpha} + W_{\beta}] \cos[(S_{\beta} - S_{\alpha})/\hbar]. \quad (10)$$

When the separate terms in Ψ do not overlap, the trajectories fall into distinct classes, there is no interference, and no problem with the interpretation. However, spatially separated classical trajectories can meet over time, and equation (10) will then be nonzero. Given that he has been assured that the mysterious equation correctly accounts for all phenomena, Maxwell concludes that phenomena displaying weird oscillatory population densities presumably exist, but he is baffled as to how to interpret them in terms familiar to him. He remarks, however, that the difference in action $|S_{\alpha} - S_{\beta}|$ between distinct trajectories must not be large compared to $\hbar$ if the oscillations of the interference term are not to render it undetectable; and he points out that if L is the dimension of the region in which the trajectories are observed, and v a characteristic velocity, the interference will average to some small quantity ϵ unless $mv \lesssim \hbar/L$, and that ϵ can be made as small as one pleases by taking the typical mass m to be sufficiently large.

Discrete degrees of freedom

Theoretical physics cannot exist in the absence of empirical knowledge about natural phenomena. Newton would therefore have been given the facts about electromagnetic phenomena that Maxwell’s equations account for when asked to unravel their meaning. In the same spirit, Maxwell is told about the phenomena that quantum mechanics accounts for. For example, he is told about the Stern–Gerlach (SG) experiment, and knows that if an atomic species has a magnetic moment, then in their motion through a magnetic field the members of this species behave as if these moments only have a discrete set of values.

In this second chapter of the fable, Maxwell is handed the Schrödinger equation for a neutral atom having a magnetic moment and moving in an inhomogeneous magnetic field. The degrees of freedom with a continuous spectrum are the position $\mathbf{q}$ in

configuration space C , while the degrees of freedom with no direct classical counterpart are in a $(2j + 1)$ -dimensional ‘internal’ Hilbert space H , spanned by sets of spherical harmonics $\{Y_\mu^j(\mathbf{u})\}$, where $\mu = j, j - 1, \dots, -j$ and j is an integer to avoid the irrelevant issue of double-values quantities. Each such set of harmonics is defined with respect to an axis $\mathbf{u}$ in three-dimensional space, and μ is the eigenvalue that specifies how $Y_\mu^j(\mathbf{u})$ behaves under rotation about $\mathbf{u}$.

The hamiltonian in this Schrödinger equation contains a time-independent term, $H_1(\mathbf{q})$, which will act as an amplifier; it is an operator both in C and H space that is nonzero only in some bounded region C_1 of configuration space. The action of $H_1(\mathbf{q})$ on the internal variables is represented by a matrix $M(\mathbf{n})$, where $\mathbf{n}$ is a constant unit vector along the direction of the field gradient. This matrix is brought to diagonal form by the particular basis $\{Y_\sigma^j(\mathbf{n})\}$, and as a consequence represents a force applied on the atoms that depends on σ , where σ , apart from a factor, is the value of the magnetic moment’s projection onto $\mathbf{n}$.

Maxwell does not find this elaboration of the formalism required by experiments of the SG variety nearly as baffling as the Schrödinger equation itself. He has mastered similar phenomena in the propagation of light through media that influence polarization, and thinks of the spherical harmonics as generalizations of polarization vectors.

Knowing of the SG experiment, Maxwell analyses a situation in which, for $t < t_1$, the spatial coordinates are described by $\psi_{\text{in}}(\mathbf{q}, t)$, a wave function that is accurately approximated by the semiclassical Ansatz of equation (6), and that corresponds to a set of classical trajectories $\{\mathbf{q}_{\text{in}}(t)\}$ forming a beam moving towards the amplifier, which it enters at $t \approx t_1$. By analogy with the treatment of optical experiments, he considers the internal state to be an arbitrary linear combination of the basis in H

$$\Xi = \sum_{\mu=-j}^j a_\mu Y_\mu^j \quad (11)$$

so that the incident beam is represented by

$$\Psi_{\text{in}}(\mathbf{q}, t) = \psi_{\text{in}}(\mathbf{q}, t) \Xi \quad (12)$$

During $t_1 < t < t_2$ this beam traverses the region C_1 where the ‘amplifier’ $H_1(\mathbf{q})$ acts; the latter is designed to be sufficiently smooth to maintain the validity of the classical approximation. Maxwell now asks what will happen to the beam as a result of having passed through the amplifier. Knowing how to compute the polarization of light as it traverses a magnetized medium, he expands equation (12) in the basis that diagonalizes $H_1(\mathbf{q})$, and shows that $\Psi_{\text{in}}(\mathbf{q}, t)$ evolves into

$$\Psi(\mathbf{q}, t) = \sum_\sigma c_\sigma \psi_\sigma(\mathbf{q}, t) Y_\sigma^j(\mathbf{n}) \quad (13)$$

where $c_\sigma = \langle Y_\sigma^j(\mathbf{n}) | \Xi \rangle$. Here each ψ_σ has a phase $S_\sigma(\mathbf{q}, t)$ determined by a JH equation with a potential that depends on σ because the effect of the amplifier is σ -dependent. The result is a set of trajectories $\{\mathbf{q}_{\text{n},\sigma}(t)\}$ that move in $(2j + 1)$ distinct directions. Each such set has a density $w_\sigma(\mathbf{q}, t) = |\psi_\sigma|^2$ satisfying a separate continuity equation, so that

$$\int d\mathbf{q} w_\sigma(\mathbf{q}, t) = C_\sigma \quad (14)$$

is a constant in the interval $t_1 < t < t_2$. But $\Psi(t) \rightarrow \Psi_{\text{in}}(t)$ as $t \rightarrow t_1$ from above, whence $\psi_\sigma(\mathbf{q}, t) \rightarrow \psi_{\text{in}}(\mathbf{q}, t_1)$, and therefore all the C_σ are equal

$$\int d\mathbf{q} w_\sigma(\mathbf{q}, t) = \int d\mathbf{q} w_{\text{in}}(\mathbf{q}, t) \quad (15)$$

By design, the region C_1 in which the amplifier acts is large enough to produce beams that are well separated for $t > t_2$, so that for such times the density is

$$w(\mathbf{q}, t) = \sum_\sigma |c_\sigma|^2 w_\sigma(\mathbf{q}, t) \quad (16)$$

Because of equation (15), Maxwell concludes that the fraction of the incident population that will end up in the beam bearing the label σ is $|c_\sigma|^2$.

Note that there are no interference terms in equation (16). As soon as the beams cease to overlap in space, interference is impossible. In any event, there is never interference between beams with differing values of σ because the Y_σ^j are orthogonal, which would be apparent to Maxwell because of the analogous property of polarization in optics.

Maxwell next considers an arrangement in which all but one of these beams, say with $\sigma = \rho$, is eliminated by another interaction term in the hamiltonian, with this filtered beam then sent through a second SG setup with some other orientation $\mathbf{k}$. By his previous argument, he concludes that that would result in a set of beams with population fractions

$$|\langle Y_\rho^j(\mathbf{n}) | Y_\sigma^j(\mathbf{k}) \rangle|^2 \quad (17)$$

All the familiar statements about the complex coefficients c_σ as probability amplitudes thus emerge from Maxwell’s analysis, though thus far only expressed as fractions of various populations that pass through some combination of filters and fields.

This then leads Maxwell to ask the fundamental question: Can a unique label ‘ σ along $\mathbf{n}$ ’ be assigned to an individual specimen that is following a trajectory in the initial set $\{\mathbf{q}_{\text{in}}(t)\}$ before this specimen enters the force field that produces the separation into the distinct trajectory sets $\{\mathbf{q}_{\text{n},\sigma}(t)\}$? If this could be done, each member of the population would then have the inherent property ‘ σ along $\mathbf{n}$ ’, which is merely revealed by the subsequent segregation into the distinct sets $\{\mathbf{q}_{\text{n},\sigma}(t)\}$. Maxwell soon realizes that this is not possible. He knows that he can expand the initial internal state Ξ in any one of the infinity of basis sets in H ; however, as the matrices $M(\mathbf{k})$ for different directions $\mathbf{k}$ do not commute, the unique choice $\{Y_\sigma^j(\mathbf{n})\}$ that identifies the quantities $|c_\sigma|^2$ as probabilities is only revealed after the beam has been separated by the amplifier oriented along the direction $\mathbf{n}$. He therefore concludes as follows:

An intrinsic property ‘ σ along $\mathbf{n}$ ’ cannot be assigned to individual specimens; all that can be said is that if a specimen is passed through the field oriented along $\mathbf{n}$, the probability that it will emerge in the population ‘ σ ’ is $|c_\sigma|^2$.

This is just the orthodox meaning of probabilities in quantum mechanics: the probability of a specific outcome as revealed by measurement.

Knowing that the magnetic moments are minute in the experiments he has been told about, Maxwell asks himself whether letting j become very large would be clarifying. His mastery of spherical harmonics allows him to show the following:

(1) If the apparatus can resolve a single beam from among the $(2j + 1)$ beams (or any set that does not grow with j), the phenomena will be similar to those just discussed when $j = O(1)$;
(2) On the other hand, if the resolution only suffices to select a fraction of the beams when $j \gg 1$, the resulting beam will have a mean magnetic moment $\bar{\mathbf{M}}(t)$ whose motion is in accord with that of classical physics, and the dispersion of the components about this mean will satisfy

$$\Delta M_x \Delta M_y \geq \hbar \lambda \Delta M_z \quad (18)$$

where $|\bar{\mathbf{M}}| = \lambda j$ (refs 3 and 4).

This inequality holds for all j , yet it is most striking to the classical mindset when $j \gg 1$. For it then states that the various components of the moment cannot all be known simultaneously to arbitrary accuracy even when $\bar{M}$ is macroscopic, though of course the uncertainties becomes negligible compared to $\bar{M}$ as $(j/\hbar) \rightarrow \infty$.

Discussion and conclusions

The fable has now served its purpose. Undoubtedly, some readers will ask why the fable does not have Maxwell arriving at a de Broglie–Bohm hidden-variable interpretation⁵, as that viewpoint is so closely tied to Hamilton–Jacobi theory. Were he to attempt this, he would be ignoring the instruction that Schrödinger’s equation “as it stands and without elaboration” does everything claimed for it. Naturally, instructions do not deter the Maxwells of this world. By the same token, Newton, on being asked to decipher Maxwell’s equations, would not have been deterred from inventing a mechanical medium whose oscillations are the ‘real’ phenomenon described by those waves propagating through nothing that first appear as solutions of the field equations. Hidden-variable embellishments of quantum mechanics have a role analogous to luminiferous-aether conceptions of electrodynamics: they provide a concrete model in terms of established and cherished preconceptions about the natural world. Only direct experimental evidence that bears on such hidden substructures can alter the terms of the debate as to whether the substructure is or is not ‘physical’. But we have such evidence: the Michelson–Morley experiment and its successors, which would have given our fictitious Newton pause; and the experiments showing that quantum mechanics ‘as it stands’ passes the tests stemming from Bell’s theorem, whereas hidden variable theories will not unless they allow action at a distance effects.

Indeed, the same combination of semiclassical and quantum mechanical descriptions can be given for experiments of the Einstein–Podolsky–Rosen type: for example, a system at rest that disintegrates into two fragments which then follow opposed classical trajectories $\{\mathbf{q}_1(t)\}$ and $\{\mathbf{q}_2(t)\}$, and a suitable correlated state in the joint internal Hilbert space $H_1 \otimes H_2$. When the widely separated fragments are passed through fields that produce distinct trajectories for the various eigenvalues (σ_1, σ_2) along the directions $(\mathbf{n}_1, \mathbf{n}_2)$, they will produce correlations that violate the Bell inequalities, with all the familiar implications that follow therefrom.

The fable showed that a classical treatment of observables with continuous spectra having a direct classical correspondence has the following merits:

- The statistical interpretation of the scalar product emerges from the Schrödinger equation for degrees of freedom in finite dimensional Hilbert spaces, as well as the uncertainty relations for observables in such spaces.
- The portions of the quantum mechanical formalism that are being used in the arguments come in through the front door, and not so surreptitiously that even the author is confused about what is assumed and what is derived⁶.
- Decoherence—the absence of interference terms in equation (16)—arises naturally as a consequence of the classical description of the classical degrees of freedom. There is no need for an external environment not included in the Schrödinger equation, nor a pyramid of devices which make laboratory demonstrations of coherence effectively impossible^{6,7}.

This argument does not, as it stands, lead to the statistical interpretation of the configuration space wave function. It does, however, make the interpretation a natural hypothesis—an extension of what has been established for observables in finite dimensional Hilbert spaces to those with a continuous spectrum. $\square$

1. Birkhoff, G. D. Quantum mechanics and asymptotic series. *Am. Math. Soc. Bull.* **39**, 681–700 (1933).
2. Schrödinger, E. *Collected Papers on Wave Mechanics* 1–2 (Blackie & Son, London, 1928).
3. Radcliffe, J. M. Some properties of coherent spin states. *J. Phys. A* **4**, 313–323 (1971).
4. Perelomov, A. M. Coherent spin states for arbitrary Lie groups. *Comm. Math. Phys.* **26**, 222–236 (1972).
5. Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics* 111–116, 127–133 (Cambridge Univ. Press, Cambridge, 1987).
6. Gottfried, K. *Quantum Mechanics* Sec. 20 (W. A. Benjamin, New York, 1966).
7. Zurek, W. H. Decoherence, einselection and the existential interpretation. *Philos. Trans. R. Soc. Lond. A* **356**, 1793–1821 (1998).
8. Bell, J. S. Against ‘measurement’. *Phys. World* Aug. 33–40 (1990).
9. Gottfried, K. Does quantum mechanics carry the seeds of its own destruction? *Phys. World* Oct. 34–40 (1991).

Acknowledgements

This is a belated, second response to John Bell’s critique⁸ of my 1966 treatment of the statistical interpretation⁶. In my first and rather unsatisfactory response⁹, written after Bell’s death but in the light of conversations with him, I interpreted his position as being based, at least in part, on the theme expressed in the opening paragraphs of this paper. I am indebted to David Mermin for advice and for posing pointed questions. This work is supported in part by the National Science Foundation.

Correspondence and requests for materials should be addressed to the author (e-mail: kg13@cornell.edu).

Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand

Jeffrey C. Boyington^{*}, Shawn A. Motyka[†], Peter Schuck[‡], Andrew G. Brooks[§] & Peter D. Sun^{*}

^{*} Structural Biology Section, Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 12441 Parklawn Drive, Rockville, Maryland 20852, USA

[†] Biochemistry, Cellular and Molecular Biology Program, Johns Hopkins University School of Medicine, 725 N. Wolfe St., Baltimore, Maryland 21205-2185, USA

[‡] Bioengineering and Physical Science Program, National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892-5766, USA

[§] Department of Microbiology and Immunology, University of Melbourne, Royal Parade, Parkville, 3052 Victoria, Australia

Target cell lysis is regulated by natural killer (NK) cell receptors that recognize class I MHC molecules. Here we report the crystal structure of the human immunoglobulin-like NK cell receptor KIR2DL2 in complex with its class I ligand HLA-Cw3 and peptide. KIR binds in a nearly orthogonal orientation across the $\alpha 1$ and $\alpha 2$ helices of Cw3 and directly contacts positions 7 and 8 of the peptide. No significant conformational changes in KIR occur on complex formation. The receptor footprint on HLA overlaps with but is distinct from that of the T-cell receptor. Charge complementarity dominates the KIR/HLA interface and mutations that disrupt interface salt bridges substantially diminish binding. Most contacts in the complex are between KIR and conserved HLA-C residues, but a hydrogen bond between Lys 44 of KIR2DL2 and Asn 80 of Cw3 confers the allotype specificity. KIR contact requires position 8 of the peptide to be a residue smaller than valine. A second KIR/HLA interface produced an ordered receptor–ligand aggregation in the crystal which may resemble receptor clustering during immune synapse formation.

Natural killer (NK) cells constitute a vital part of the innate immune system^{1,2}. Like cytotoxic T-cells, NK cells express surface receptors that interact with polymorphic class I MHC molecules and regulate cell lysis. In contrast to T-cell receptor (TCR)/MHC-mediated cellular activation, recognition of class I molecules by NK cells can result in either target cell lysis or the inhibition of lysis depending on whether the receptor contains a charged transmembrane residue that interacts with the immunoreceptor tyrosine-based activation (ITAM) motif containing DAP-12 (ref. 3) or with a cytoplasmic immunoreceptor tyrosine-based inhibitory (ITIM) motif. Furthermore, MHC recognition by NK cell receptors is less allele specific and less peptide dependent. Of particular interest are inhibitory NK cell receptors that protect target cells bearing certain class I MHC allotypes from NK-mediated lysis⁴. Cells that have lost class I MHC expression such as certain tumour or virus-infected cells will not be engaged by inhibitory receptors and consequently may become susceptible to NK cell lysis⁵. The two structurally distinct superfamilies of NK cell receptors are the immunoglobulin-like and the C-type lectin-like receptors (CTLR). To date, the crystal structures of three killer cell immunoglobulin-like (KIR) receptors, KIR2DL1, 2DL2 and 2DL3, and two CTLR receptors, human CD94 and murine Ly49A in complex with H-2D^d, have been determined^{6–10}.

Previous studies indicated that class I HLA recognition by KIR depends on residues 77 and 80 of the class I HLA heavy chain and peptide residues 7 and 8 (ref. 11). The inhibitory receptor KIR2DL2 interacts with the Ser 77 and Asn 80 bearing HLA-Cw1, 3, 7 and 8 allotypes; however, other residues on the HLA molecule must also contribute to the recognition by KIR, because many HLA-B alleles with the same Ser 77 and Asn 80 as HLA-Cw3 are not recognized by KIR2DL2. Receptor mutagenesis has shown that residues 44, 45 and 70 influence class I HLA binding^{11–13}. These studies and the crystal structure of KIR2DL2 have suggested a putative ligand-binding region on the receptor⁷. We now report the crystal structure of KIR2DL2 bound to HLA-Cw3 and provide a molecular mechanism for the allotypic specificity and peptide recognition by KIR molecules on human NK cells. This also reveals a different binding

mode for KIR2DL2 from that of the murine C-type lectin-like receptor Ly49 with its class I ligand H-2D^d despite their functional similarity¹⁰.

Overall structure of the complex

The structure of KIR2DL2 in complex with HLA-Cw3 and the peptide GAVDPLAL (GAV) was determined by molecular replacement methods and refined to 3.0 Å resolution. The final *R* factors are 23.1% and 29.4% for *R*_{cryst} and *R*_{free}, respectively (Table 1). Each asymmetric unit contains two KIR (KIR A and KIR B) and one HLA-Cw3 molecule. The electron density is continuous in the final

Table 1 Statistics for data collection and refinement

Data collection	
Space group	P2 ₁ 2 ₁ 2 ₁
Cell dimensions	<i>a</i> = 68.5, <i>b</i> = 90.3, <i>c</i> = 207.3 Å
Resolution limit	3.0 Å
Unique reflections	24,517 (1976)*
Redundancy	4.2 (2.2)*
Completeness (%)	92.1 (75.4)*
<i>R</i> _{sym} (%)†	7.3 (32.7)*
<i>I</i> / <i>σ</i> _{<i>I</i>}	15.6 (2.0)*
Refinement (10.0–3.0 Å, <i>F</i> / <i>σ</i> _{<i>F</i>} ≥ 1.0)	
No. reflections	21,717
No. nonhydrogen atoms	6411
Solvent molecules	185
<i>R</i> _{cryst} (%)	23.1 (23.4)‡
<i>R</i> _{free} (%)§	29.4 (29.9)‡
Mean B-factor (Å ²)	63.5
Wilson plot B-factor (Å ²)	63.7
r.m.s. deviation from ideality	
Bonds (Å)	0.008
Angles (°)	1.47

*Values for the highest resolution shell (3.11–3.0 Å).

†*R*_{sym} = 100 × Σ(*I*_{*h*} − *I*_{*h*})/Σ*I*_{*h*}, where *I*_{*h*} is the mean intensity of multiple measurements of symmetry related reflections.

‡Calculated using all *F* > 0.

§Five per cent of the reflections were used as a test set for calculating *R*_{free}.

||Mean B-factors (Å²) for individual domains are HLA-Cw3; α1α2, 58.2; α3, 68.9; β2m, 70.1; GAV: 47.4; KIR A: D1, 50.5; D2, 60.5; KIR B: D1, 71.1; D2, 73.5; H₂O: 52.4.

$2F_o - F_c$ map throughout the complex except for four KIR surface loops (residues 143–145 of KIR A and residues 58, 121 and 158–159 of KIR B) located in regions away from the HLA interface. Only KIR A interacts with HLA-Cw3 (Fig. 1). KIR B is situated on the distal end of the D2 domain of KIR A, providing lattice contacts between symmetry related receptors in the crystal. Thus the stoichiometry of KIR:HLA binding is 1:1 in the crystal. This is also supported by analytical equilibrium centrifugation experiments in which the dissociation constant (K_d) for KIR2DL2 and HLA-Cw3 association was measured as $17 \mu\text{M}$ and sedimentation curves were best described by a 1:1 interaction model (data not shown).

KIR A interacts with HLA-Cw3 through surface loops near its interdomain hinge region. The binding site for KIR is located

toward the carboxy-terminal end of the peptide and the corresponding region of the $\alpha 1$ and $\alpha 2$ helices. The orientation of the KIR is such that its amino-terminal D1 domain interacts with polymorphic regions of the $\alpha 1$ helix, residues 69–84, and its D2 domain interacts with more conserved regions of the $\alpha 2$ helix, residues 145–151. This produces a nearly orthogonal (88°) docking of KIR to its class I ligand and permits direct contact between KIR and residues 7 and 8 of the GAV peptide.

The structure of KIR2DL2

KIR A is nearly identical to KIR B and to the structure of ligand free KIR with root-mean-square (r.m.s.) differences of $1.04 \text{ \AA}$ (194 C α pairs) and $0.97 \text{ \AA}$ (191 C α pairs), respectively⁷. The electron density for KIR A is better than for the unligated KIR B. The hinge angle between the D1 and D2 domains of KIR A, KIR B and free KIR are 81° , 85° and 84° , respectively⁷. Binding of KIR A to HLA-Cw3 results in the displacement of CC' loop residues Lys 44 and Phe 45 by $2.6 \text{ \AA}$ relative to KIR B and a small change in the interdomain hinge angle. No other significant conformational differences were observed between the HLA-Cw3 contacting loops of KIR A and the corresponding loops of KIR B and free KIR. KIR B is positioned atop the D2 domain of KIR A, burying $710 \text{ \AA}^2$ of surface area (Fig. 1). The biological relevance of this KIR A/KIR B interaction is, however, not clear.

The structure of HLA-Cw3

The overall structure of HLA-Cw0304 displays the salient features of other classical class I MHC molecules, except for subtle differences in the relative orientations of the $\alpha 1\alpha 2$, $\alpha 3$ and $\beta 2m$ domains. Hinge angles between the $\alpha 1\alpha 2$ and $\alpha 3$ domains in class I molecules vary from 55 to 84° , with most between 76 and 80° . The structure of HLA-Cw3 has a nearly orthogonal 87° hinge angle. The $\alpha 1\alpha 2$ domains of HLA-Cw3 and HLA-Cw4 share r.m.s. differences of $0.82 \text{ \AA}$ (181 C α pairs). The most visible differences are in three variable surface loops of the $\alpha 1\alpha 2$ domain (14–20, 39–44 and 104–107) and in the width of the peptide-binding groove¹⁴. The peptide-binding cleft of HLA-Cw3 near the regions of $\alpha 1$ residues 70–77 and $\alpha 2$ residues 144–152 is $\sim 2.5 \text{ \AA}$ narrower than that of Cw4, but is

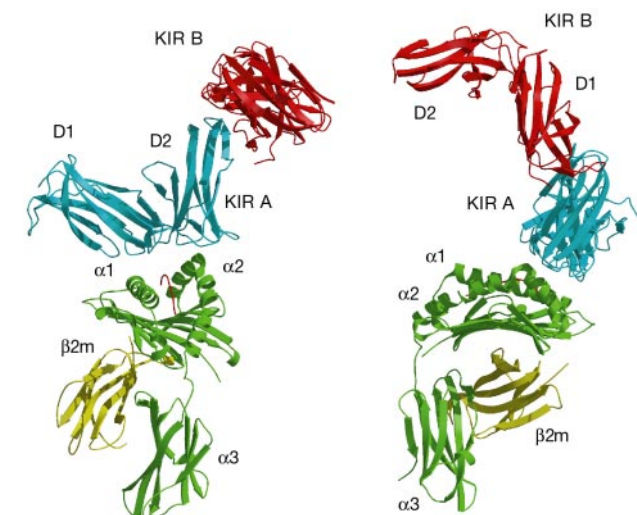


Figure 1 Ribbon drawing showing two views of HLA-Cw3 bound to KIR2DL2. The $\beta 2m$ domain, HLA-Cw3 heavy chain and peptide are yellow, green and magenta, respectively. The KIR molecule bound to HLA-Cw3 and the second KIR molecule are blue and red, respectively. Right view is rotated $\sim 90^\circ$ from left view along the vertical axis. All figures were created using the programs MOLSCRIPT⁴¹, RASTER3D⁴² and GRASP⁴³.

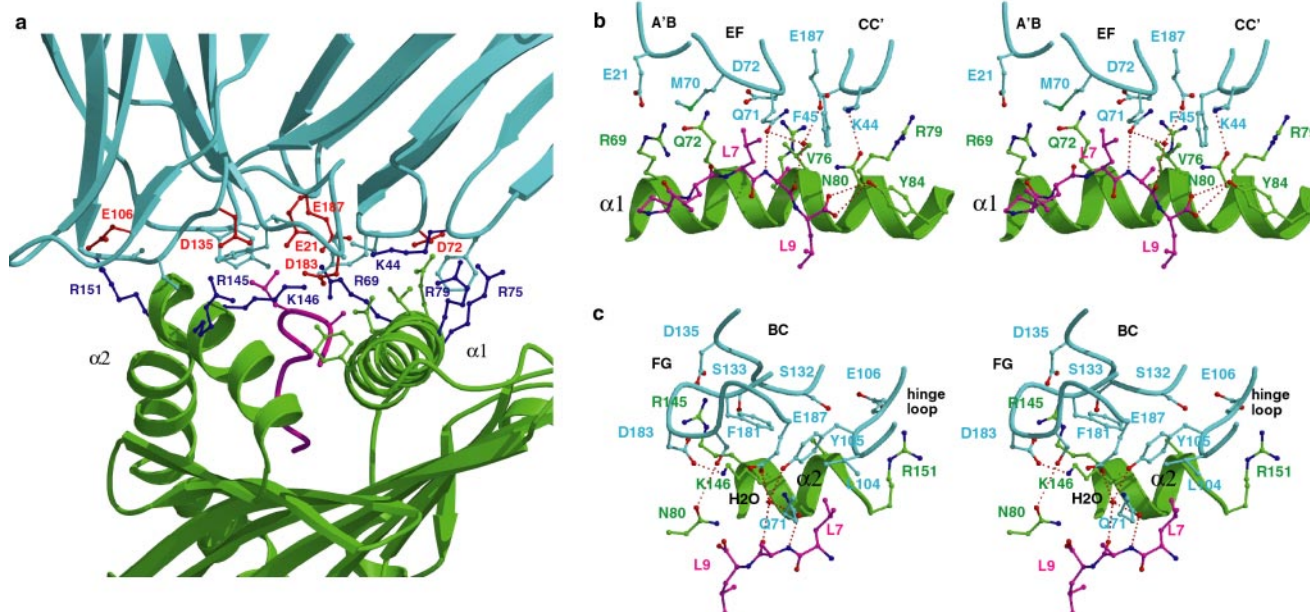


Figure 2 The KIR2DL2/HLA-Cw3 interface. **a**, Charge complementarity at the interface. Basic residues are dark blue, acidic residues are red, and the remaining residues are coloured by molecule. **b,c**, Stereo view of the interaction between domain D1 of KIR, the GAV peptide and the $\alpha 1$ helix of HLA-Cw3 (**b**), and domain D2 of KIR, the hinge loop, the

GAV peptide and the $\alpha 2$ helix of HLA-Cw3 (**c**). KIR is shown in light blue, the GAV peptide in purple and HLA-Cw3 in green. Selected hydrogen bonds are represented by dotted lines.

comparable to that of most other human and mouse class I MHC molecules.

The KIR/HLA interface and its charge complementarity

KIR2DL2 and HLA-Cw3 complex formation buries a total of 1,562 Å², with a predominance of charged and hydrophilic residues at the interface forming an extensive array of hydrogen bonds and salt bridges. The interface is relatively flat with poor shape complementarity (mean shape correlation statistic¹⁵ of 0.58), comparable to those of TCR/MHC, and the CD2–CD58 adhesion complex, but less than those of antibody–antigen interfaces^{16,17}. However, the KIR/HLA interface possesses strong charge complementarity. The interface region on KIR contains one basic (Lys 44) and six acidic residues (Glu 21, Asp 72, Glu 106, Asp 135, Asp 183 and Glu 187), whereas the complementary site on HLA-Cw3 has six basic (Arg 69, Arg 75, Arg 79, Arg 145, Lys 146 and Arg 151) and no acidic residues (Fig. 2a). This results in the formation of four salt bridges between KIR and HLA-Cw3 (Glu 21–Arg 69, Glu 106–Arg 151, Asp 135–Arg 145 and Asp 183–Lys 146). In addition there are eight hydrogen bonds at the interface (Table 2). The charged side chains are evenly distributed over both the KIR and HLA-Cw3 binding sites rather than forming clusters. Each of the six binding loops on KIR has one or more charged residues and the α1 and α2 helices on HLA-Cw3 each have three basic residues in the binding region.

To assess the importance of these interface charge–charge interactions in KIR/HLA recognition, three of the four salt bridges were mutated individually on KIR2DL2. The effect of these single salt-bridge mutations, E106A, D135H and D183A, on HLA-Cw3 binding was measured using surface plasmon resonance (SPR) techniques. As controls, an R33A mutation, well outside the KIR/HLA interface, and a K44M mutation were also created. While the *K_d* of R33A is essentially identical to that of wild-type KIR2DL2, the HLA binding affinity of E106A is reduced by 6-fold and affinities of D135H and D183A are each 20-fold lower than that of the wild type (Table 3). The K44M mutation resulted in a near complete loss of HLA-Cw3 binding, consistent with previous results obtained from cell-surface binding assays¹². The necessity of these salt bridges for successful KIR/HLA recognition emphasizes the crucial role of charge in

HLA recognition and reveals their low tolerance to interface mutations. This suggests a mechanism for the relatively low-affinity KIR receptors to achieve their ligand specificity by requiring a high binding energy threshold for recognition. In contrast, the haematopoietic receptors are more tolerant of mutations.

In total, six loops from KIR interact with the HLA-Cw3/GAV molecule. They include loops A'B (20–23, connecting β-strands A' and B), CC' (43–46) and EF (67–74) of the D1 domain, the hinge loop (103–108), and loops BC (130–135) and FG (182–184) of the D2 domain. Loops A'B, CC' and EF of the D1 domain contact the α1 helix of HLA-Cw3 and the backbone of the GAV peptide with Glu 21 of KIR forming a salt bridge to Arg 69 of HLA-Cw3 (Fig. 2b). Lys 44 from the KIR CC' loop forms a hydrogen bond with Asn 80 of HLA-Cw3. Both residues have been shown to be pivotal in allotypic recognition^{12,18}. A direct hydrogen bond is formed between Gln 71 (Oε1) of the EF loop in KIR and the amide nitrogen of residue Ala 8 in the GAV peptide. This results in close contact between the peptide Ala 8 Cβ atom and the Cδ atom of Gln 71. One of two clusters of nonpolar interactions at the KIR/HLA interface is centred around Val 76 on the α1 helix of HLA-Cw3 and Phe 45 on the CC' loops of KIR. Included in this cluster are the aliphatic portions of Arg 69, Arg 75 and Arg 79 of HLA-Cw3 and Lys 44, Met 70, Gln 71 and Asp 72 of KIR. A simple mutation of Phe 45 to tyrosine (naturally present in the non-inhibitory receptor, KIR2DS2) significantly reduced the affinity of KIR for HLA-Cw3 (ref. 11). The tip of the Phe 45 is only 3.2 Å from the Cβ of Arg 79 in the α1 helix of HLA-Cw3, leaving little space to accommodate an additional hydroxyl group.

The hinge loop and the BC and FG loops of the D2 domain of KIR each form a salt bridge with basic residues (Arg 151, Arg 145 and Lys 146, respectively) on the α2 helix of HLA-Cw3 (Table 2, Fig. 2c). Tyr 105 and Phe 181 together with HLA-Cw3 residues Ala 150 and Lys 146 (aliphatic portion) and Leu 7 of GAV form the second hydrophobic cluster at the interface. In addition, Gln 71, Tyr 105 and Glu 187 of KIR form water-mediated (HOH 38) hydrogen bonds with the carbonyl oxygen of Ala 8 of GAV. The mutation of Tyr 105 to alanine resulted in a 20-fold decrease in HLA-Cw3 binding affinity (Table 3). This mutation presumably destabilized the interface hydrophobic packing. Notably, KIR2DL2 uses the

Table 2 Interactions between KIR2DL2 and HLA-Cw3/GAV

KIR2DL2		HLA-Cw3/GAV		Distance (Å)
Hydrogen bonds* and salt bridges				
Glu 21	Oε1	Arg 69	NH2	3.2
Glu 21	Oε2	Arg 69	NH2	3.2
Lys 44	Nζ	Asn 80	Oδ1	3.3
	O	Arg 79	Nε	3.4
Gln 71	Oε1	Ala 8†	N	3.4
Asp 72	Oδ2	Gln 72	Nε2	3.4
Glu 106	N	Ala 149	O	2.9
	Oε2	Arg 151	NH2	2.7
Ser 133	O	Arg 145	NH2	2.8
	Oγ	Arg 145	Nε	2.9
	Oγ	Arg 145	NH2	3.2
Asp 135	Oδ2	Arg 145	NH2	2.6
Asp 183	Oδ2	Lys 146	Nζ	2.5
Hydrophobic contact†				
Phe 45		Arg 75(6), Val 76(5), Arg 79(4)		
Met 70		Arg 69(1)		
Gln 71		Val 76(1), Ala 8†(1)		
Asp 72		Val 76(1)		
Leu 104		Leu 7†(2)		
Tyr 105		Ala 150(1), Lys 146(3), Leu 7†(2)		
Ser 132		Ala 149(1)		
Phe 181		Lys 146(8)		
Asp 183		Lys 146(1)		

* Acceptor–donor distances 2.5–3.4 Å.

† Carbon–carbon contacts ≤4.0 Å. Parentheses indicate the number of contacts between a particular residue pair.

‡ GAV peptide residue.

Table 3 Peptide and receptor mutation effects in KIR2DL2/HLA-Cw3 association

Effects of peptide variation in KIR/HLA binding			
Peptide		<i>K_d</i> (μM)	w6/32 binding (%)
GAVDPLAL	(GAV)	9.5	100
-----S-	(GAV_S)	42.3	147
-----V-	(GAV_V)	525	130
-----Y-	(GAV_Y)	>600	130
-----K-	(GAV_K)	>600	139
AAADAAAL	(AAA)	48.5	149
TAMDVVYAL	(TAM)	38	138
QAISPRTL	(QAI)	74	38
HLA-E		>600	84
Effects of amino acid substitutions in KIR/HLA association			
KIR mutant		<i>K_d</i> (μM)	
Wild type		28	
R33A		30	
K44M		>400	
Y105A		>400	
E106A		185	
D135H		>400	
D183A		>400	

For the peptide variation measurements, soluble KIR2DL2 receptor and w6/32 were immobilized on CM5 sensor chips. Binding was measured with a serial dilution of HLA-Cw3 from 0.3 to 40 μM and the dissociation constant *K_d* was determined by a steady-state model for all peptides. The w6/32 binding, listed as the percentage of the HLA-Cw3/GAV binding, was determined using a 9.3 μM concentration of various HLA peptide complexes. The measurements on mutational KIR/HLA association were done similarly with immobilized KIR mutants and a serial dilution for HLA-Cw3/GAV as the analyte.

same six binding loops as do the topologically similar haematopoietic receptors. Unlike the haematopoietic receptors, which interact with their ligands as dimers and have larger hinge angles of 90° or more, the KIR receptor binds to HLA-Cw3 in a 1:1 stoichiometry.

The interface area between KIR and HLA-Cw3 (1,562 Å²) is similar to that observed in TCR/MHC complexes (1,700–1,880 Å²). Furthermore, the two receptors share similar docking orientations (Fig. 3a). There are, however, distinct differences in their recognition of class I MHC antigens. First, the footprint of KIR on HLA-Cw3 is distinct from that of TCR (Fig. 3b, c). The KIR-binding site is centred near the C-terminal P7 and P8 positions of the peptide whereas the known TCR/MHC interfaces are centred near the middle P4–P6 positions of their peptides (Fig. 3b). Their footprints, do, however, overlap partially, suggesting that there is a mutually exclusive binding between KIR and TCR on the same HLA molecule. Second, the ligand-binding domains of KIR (D1 and D2) are two tandem Ig-like domains, whereas the analogous Vα and Vβ domains of TCR are on different chains. Last, unlike TCR/MHC interactions, the KIR/MHC interactions appear to be dominated by charge–charge interactions, that more closely resemble the interaction between the adhesion molecules CD2 and CD58 (ref. 17).

Peptide binding and a potential induced-fit mechanism

The nonamer GAV peptide is derived from the human importin-α-1 subunit (residues 204–212). Its binding to HLA-Cw3 displays characteristics of the classical MHC class I peptide-binding mode in which both termini are securely anchored by a conserved pattern of hydrogen bonds and the middle residues form an arch above the floor of the cleft (Fig. 4a). Side chains from three of the nine residues, P4, P7 and P8, point out of the binding cleft. Of the other six residues, five of them, P1, P2, P3, P5 and P9, are at least 90% buried and P6 is 72% buried. The dimorphic positions Ser 77 and Asn 80 hydrogen bond with the P9 amide nitrogen and carboxylate, respectively. One feature notably missing in HLA-Cw3 as compared with HLA-Cw4 and other class I molecules is a conserved hydrogen bond between Lys 146 and the P9 carboxylate. Instead, Lys 146 of HLA-Cw3 forms a salt bridge with KIR residue Asp 183. A tyrosine in place of serine at position 9 of HLA-Cw3 significantly restricts the size of the P2 pocket, rendering it more similar to HLA-A2 than to HLA-Cw4. As a result, the P2 pocket favours alanine in HLA-Cw3 (refs 19, 20) and tyrosine in HLA-Cw4 (ref. 14).

The most unusual peptide conformation occurs at the position 7 of the peptide. Although P7 in most other known class I MHC nonamer structures has a slightly downward and mainly buried

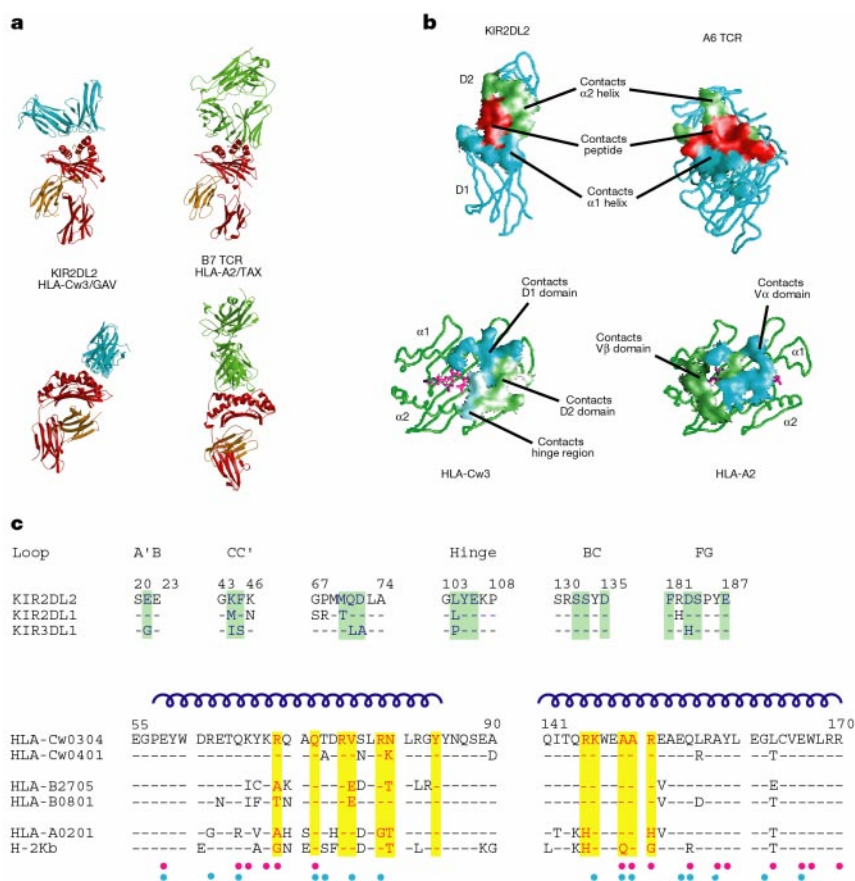


Figure 3 Comparison of KIR/HLA-Cw3 and TCR/MHC interactions. **a**, Docking orientations of KIR2DL2 (left) and A6 TCR (right, PDB accession code 1BD2) onto their class I MHC ligands. The bottom view is orthogonal to the top view. **b**, Worm representations of KIR2DL2 (top left), HLA-Cw3 (bottom left), the A6 TCR (top right, PDB accession code 1A07) and HLA-A2 (bottom right). For greater clarity, the α3 and β2m domains of HLA-Cw3 and HLA-A2 are not shown. The binding regions (intermolecular contact ≤4 Å) are depicted by molecular surface representations, coloured according to contact regions as labelled. GAV and Tax peptide residues that do not interact with KIR2DL2 or the A6 TCR, respectively, are shown as purple stick models. **c**, Structure-

based sequence alignments of the binding regions of KIR2DL2 and HLA-Cw3 with selected KIR receptors and class I MHCs. The sequence of KIR2DL3 is not shown because all six binding loops are identical to those of KIR2DL2. HLA-B2705 and HLA-B0801 are representative of Bw4 and Bw6 serotypes, respectively. Conserved residues are indicated by dashes. Helices are denoted by blue spirals. KIR residues that align with KIR2DL2 contact residues are blue on a green background. MHC residues that align with HLA-Cw3 contact residues are red on a yellow background. Red and blue dots at the bottom indicate KIR residues that interact with TCR in the HLA-A2/Tax/A6 TCR and H-2K^b/dEV8/2C TCR complexes, respectively.

conformation, Leu 7 of GAV is oriented upward, away from HLA-Cw3 and contacts KIR2DL2. Surface accessibility calculations indicate that the P7 position is normally 60–90% buried in the MHC peptide-binding cleft (Fig. 4b). In comparison, Leu 7 of GAV is only 38% buried in the HLA-Cw3 peptide-binding cleft despite being a hydrophobic residue. In marked contrast, the P7 position of HLA-Cw4 is completely buried¹⁴. These results indicate that the P7 of GAV may be hyperexposed relative to other class I MHC peptides at this position. Possibly, the interaction of KIR2DL2 with HLA-Cw3 induces a conformational change at the P7 position of GAV, resulting in the exposure of Leu 7 to KIR. It should be noted, however, that the peptide conformation of Leu 7 in the absence of KIR2DL2, is not known. In the complex, the side chain of leucine at P7 packs against Leu 104 and Tyr 105 of KIR, but this packing leaves sufficient space to accommodate different amino acids, such as tyrosine in the TAM peptide (Table 3). A similar change in peptide conformation involving position P7 was observed when the A6 TCR/HLA-A2/Tax complex was compared with the structure of HLA-A2/Tax alone²¹.

KIR binding imposes the peptide preference at position P8

The peptide contribution to the HLA-Cw4 allotype recognition by KIR2DL1 is limited to the P7 and P8 positions²²; however, the peptide preference for KIR2DL2/HLA-Cw3 recognition is less defined. In the complex structure, in addition to the hydrogen bond between the amide nitrogen of Ala 8 and Gln 71, Ala 8 is also surrounded by KIR residues Lys 44, Ser 184 and Asp 187. The particularly close proximity of Gln 71 substantially restricts the residue size at the P8 position, leaving little room to accommodate a side chain larger than valine or threonine. To examine further the P8 position size requirement, we used SPR techniques to measure the affinity of several variations of the GAV peptide with serine,

valine, lysine and tyrosine substitutions at position P8 (Table 3). Analysis of the binding of KIR2DL2 to HLA-Cw3 molecules refolded with these variant peptides revealed rapid kinetics ($k_{on} > 10^4 \text{ M}^{-1} \text{ s}^{-1}$ and $k_{off} > 0.5 \text{ s}^{-1}$), consistent with previous SPR studies of KIR2D/HLA-C recognition^{23,24}. In general, KIR/HLA-C interaction has kinetics that are characteristic of low-affinity adhesion molecules, such as CD2 and CD58 (ref. 25), but that are faster than the kinetics of TCR/MHC interaction²⁴. Whereas all peptide variants of HLA-Cw3 bind to the class I specific antibody W6/32 with nearly equal affinity, their binding affinities for KIR2DL2 differ markedly (Table 3). Small amino acids like alanine and serine at the P8 position bound well to the receptor, whereas larger amino acids such as tyrosine and lysine essentially abolished binding to KIR. The binding of HLA-Cw3 with a motif peptide AAADAAAAL to KIR2DL2 was slightly lower than those of the alanine and serine variants but much better than the valine and tyrosine variants, suggesting that although other positions contribute to KIR recognition they are less critical than the P8 position.

Allotype specificity of KIR/HLA recognition

A primary difference between the KIRs and TCRs in MHC recognition is in the nature of peptide and MHC specificity. T-cell receptors are often restricted to a particular MHC allele and its peptide, but KIR recognition of class I MHC is mainly allotype specific. Peptide tolerance of KIR recognition is evident from the fact that only the P7 and P8 positions of the peptide contact the receptor, and P4, P5 and P6 positions make no contact with the receptor. KIR2DL2 buries ~36% of the exposed peptide surface of HLA-Cw3/GAV. In comparison, ~80% of the HLA-A2/Tax and the H-2K^b/dEV8 peptide surfaces are buried upon binding to their respective A6 and 2C TCRs (Fig. 3b). The low contribution of peptide to KIR2DL2/HLA-Cw3 interaction relative to that of TCR/MHC class I interaction renders KIR less sensitive to the identity of the peptides.

The basis for the MHC allotypic recognition of KIR is reflected in the composition of the KIR/HLA interface. Of the 12 HLA-Cw3 residues in the interface with KIR, 11 are invariant in all HLA-C alleles (Fig. 3c). In contrast, only 8 out of 16 residues of HLA-A2 in the A6 TCR interface are invariant in HLA-A alleles. Thus, KIR recognition of HLA-C alleles relies more on conserved residues relative to TCRs that interact with more polymorphic residues. Of particular interest is residue 80 of HLA-C which varies between asparagine in KIR2DL2-recognizing HLA-Cw1, 3, 7 and 8, and lysine in KIR2DL2-recognizing HLA-Cw2, 4, 5, 6 and 15 (ref. 26). Of the 16 KIR residues that contribute to the KIR/HLA interface, all are invariant in KIR2DL2 and 2DL3, and 14 are identical in 2DL1 (Fig. 3c). Only two residues differ between KIR2DL2 and 2DL1—Lys 44 and Met 70 in 2DL2, and Met 44 and Thr 70 in 2DL1. The high degree of sequence conservation of the interface residues on both KIR and HLA-C suggests a common binding mode for the KIR2DL1, 2DL2 and 2DL3 receptors. Residue 44 of KIR and residue 80 of HLA-C have been implicated as critical residues in defining allotype specificity^{12,27}. Indeed, Lys 44 of KIR2DL2 forms a hydrogen

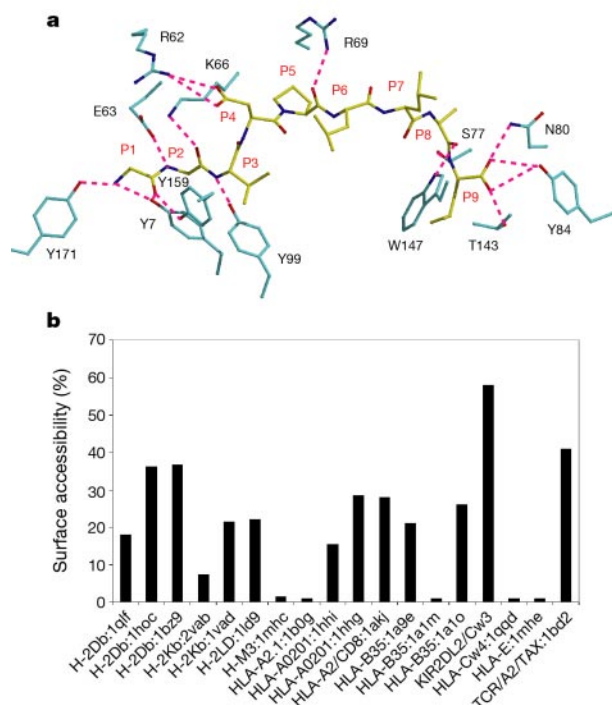


Figure 4 Coordination of the GAV peptide. **a**, Stick model showing the hydrogen-bonding pattern (dotted lines) between HLA-Cw3 (cyan) and the GAV peptide (yellow).

b, Percentage peptide surface accessibility at peptide position P7. The survey includes nonamer peptide–MHC class I molecule structures deposited in the Protein Data Bank (codes are given along the abscissa). In the case of KIR/MHC and TCR/MHC complexes, the calculation was done without the presence of the receptors.

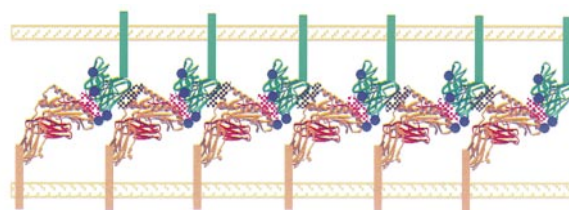


Figure 5 KIR/HLA aggregation. The complex of KIR2DL2 and HLA-Cw3 forms a regular oligomeric aggregate in the crystal lattice. KIR and HLA-Cw3 are shown in green and orange, respectively. The functional and oligomeric KIR/HLA-Cw3 interfaces are highlighted in pink and grey, respectively. The predicted glycosylation sites (highlighted by blue dots) are at residues 63, 157 and 190 for KIR, and at 86 for HLA-Cw3.

bond with Asn 80 of HLA-Cw3 in the complex structure. This hydrogen bond would have been lost when either Lys 44 of KIR2DL2 was replaced with methionine, as in KIR2DL1, or Asn 80 was replaced with lysine, as in the HLA-Cw2, 4, 5, 6 and 15 allotypes. Indeed, the K44M mutation of KIR2DL2 causes complete loss of the HLA-Cw3 binding (Table 3). Thus, one hydrogen bond preferentially stabilizes the KIR2DL2/HLA-Cw3 pair over both 2DL1/HLA-Cw3 and 2DL2/HLA-Cw4 pairs and is sufficient in determining the receptor–ligand specificity. The result of K44M mutation further supports the notion that the threshold of KIR/HLA recognition is high and the recognition is intolerant to interface mutations.

The three-domain receptor, KIR3DL1, binds to HLA-B alleles, specifically the Bw4 serotype. A comparison with HLA-B alleles reveals that all but three of the interface residues (Arg 69, Val 76 and Asn 80) are conserved across HLA-B and HLA-C loci (Fig. 3c). In HLA-B alleles, position 69 is an alanine or threonine, position 76 is a glutamic acid and position 80 is an asparagine, threonine or isoleucine. These three residues may determine locus specificity between HLA-B and HLA-C alleles. Indeed, HLA-B46, an unusual B allele that has HLA-Cw1 residues in positions 66–76, inhibits NK cells with HLA-Cw1 specificity²⁸. Nine of the sixteen KIR2DL2 interface residues are conserved in the second and third extracellular domains of KIR3DL1 (Fig. 3c). Moreover, peptide position P7 and P8 are also important for NK cell recognition of HLA-B2705 (ref. 29). Although the D0 domain of KIR3DL1 is also required for ligand recognition³⁰, the above observations suggest that KIR3DL1/HLA-Bw4 interaction is similar to that of KIR2DL2/HLA-Cw3.

The structure of a murine C-type lectin-like NK receptor Ly49A in complex with its class I MHC ligand, H-2D^d has been determined¹⁰. Although both KIR2DL2/HLA-Cw3 and Ly49A/H-2D^d recognition are dominated by charge–charge interactions, they bind to opposite ends of the class I molecule with direct peptide contact observed with KIR, but not with Ly49A. The binding of class I by KIR is, in fact, more similar to that of the TCR than to that of Ly49A. The differences between TCR, KIR and Ly49A in their recognition of class I MHC molecules illustrate both the multiplicity and complexity of the immune system.

Ligand-induced receptor aggregation

This structure and previous solution studies²⁴ suggest a 1:1 stoichiometry for KIR/HLA interaction. It is not yet clear how the signal transduction machinery detects the engagement of KIR by class I, but the formation of ligand-induced higher order oligomers is a likely mechanism. Within the KIR/HLA complex crystal and apart from the binding interface, the first KIR molecule (KIR A) also makes an additional contact with a translationally related HLA-Cw3 in a peptide-independent manner. This interaction forms an oligomeric KIR/HLA aggregate, linking adjacent pairs of complexes in the same orientation that also maintains the 1:1 molar ratio between KIR and HLA (Fig. 5). The contact surface is formed between the B and E β -strands of the KIR D2 domain and the C-terminal end of the α 2 helix of HLA-Cw3 (Fig. 5). This interface buries 528 Å² of surface area and is characterized by mostly van der Waals interactions. The putative glycosylation sites on both KIR2DL2 and HLA-Cw3 are outside this oligomeric KIR/HLA interface. It is possible that this form of receptor–ligand oligomerization resembles the receptor clustering on the surface of NK cells during immune synapse formation. However, further studies are needed to address the biological relevance of this oligomeric form, particularly its implication for receptor signalling.

Shared recognition between KIR and TCR

As part of the immunosurveillance system, classical class I MHC molecules present foreign peptides to cytotoxic T-cells through TCRs, thereby eliciting T-cell responses. The very same MHC

molecules can also present self peptides to NK cells through inhibitory KIR receptors in the absence of foreign peptides and thereby confer protection of healthy cells against NK-directed cytotoxicity. Whether KIR receptors preferentially recognize a pool of peptides distinct from those recognized by TCR is not clear. In the case of inhibitory receptors, however, it may be beneficial if such recognition of foreign peptides by KIR were less stable than the self-recognition by KIR. CD8 positive T cells that are reactive against HLA-Cw3 containing the HIV gag peptide QAISPTL have been identified³¹. This octameric peptide has a threonine at the P Ω -1 position and a fivefold reduced affinity for KIR2DL2 relative to GAV when complexed with HLA-Cw3 (Table 3). An HLA-Cw4 restricted T-cell epitope peptide has been shown not to be recognized by KIR2DL1 (ref. 22); it is therefore possible that KIR receptors may be optimized to bind to MHC with self peptides. The situation may, however, be different in murine system where the Ly49 family of receptors appears less sensitive to the peptide identity^{32,33}.

This work provides the first structural insight to KIR/HLA recognition, but many questions remain unanswered. What is the structural role of zinc ion in KIR receptor function and what is the structural organization of the NK cell immune synapse^{34,35}? Does KIR3D recognition of HLA-B alleles share the same structural features predicted from the KIR2DL2/HLA-Cw3 interface? Further structural work is needed to address these questions and to allow us to understand better the machinery of innate immune recognition. □

Methods

Protein expression, purification and crystallization

KIR2DL2 was expressed and purified as described⁷. The heavy chain of HLA-Cw304 (1–278) and β 2m (0–99) were expressed separately as *Escherichia coli* inclusion bodies and then reconstituted *in vitro* together with the peptide GAVDPLAL, purified on a Source Q anion exchange column and a Superdex 200 gel-filtration column. The complex of KIR2DL2 and HLA-Cw3 was prepared by mixing both components in a 1:1 molar ratio. Crystals were grown at room temperature by hanging drop vapour diffusion methods to 0.05 $\times$ 0.1 $\times$ 0.4 mm in size using 6% PEG 20,000, 50 mM CaCl₂ and 50 mM sodium cacodylate at pH 6.5 and 10 mg ml^{−1} of the complex.

Structure determination

All data were collected at −180 °C using an ADSC Quantum IV CCD detector on the X9B beam line of the National Synchrotron Light Source (NSLS) at the Brookhaven National Laboratory and processed using HKL2000 (ref. 36). Before freezing, crystals were soaked briefly in precipitant solution containing 25% glycerol. The crystals diffract to 3.0 Å resolution and belong to the orthorhombic space group P2₁2₁2₁ with unit-cell dimensions of $a = 68.5$, $b = 90.3$ and $c = 207.3$ Å. There are one HLA-Cw3/GAV peptide and two KIR2DL2 molecules in the asymmetric unit.

The structures of HLA-Cw3 and both KIR2DL2 molecules were determined by molecular replacement using the program AmoRe³⁷. The polyaniline version of HLA-A2 (PDB accession number 1B0G) was used as the model in rotation and translation searches using 10–3.5 Å data. This yielded a clear solution with a correlation coefficient of 40.0% and an R factor of 53.8%. After rigid body refinement of individual domains using the program CNS³⁸, most side chains of HLA-Cw3 had clear electron density into which all but 40 side chains were built. The first KIR2DL2 molecule was identified by a search using a polyaniline model of the unbound KIR2DL2 (PDB accession code 2DL2) as a model and with the HLA-Cw3 position fixed. The solution had a correlation coefficient of 44.1% and an R factor of 51.0%. After preliminary refinement in CNS, a second KIR molecule was visible in the $F_o - F_c$ map contoured to 2.0 σ and its position was defined by further searches using the fixed positions of HLA-Cw3 and the first KIR. This yielded a final solution with a correlation coefficient of 66.6% and an R factor of 40.1%.

Positional and grouped B -factor refinement was carried out using the maximum likelihood method of CNS including a bulk-solvent correction and anisotropic B -factor scaling of F_o in conjunction with remodelling using the program O (ref. 39). Reflections from 10.0 to 3.0 Å with $F_o \geq 1.0\sigma_{F_o}$ were used for refinement, excluding 5% that were used for R_{free} calculations. Non-crystallographic symmetry (NCS) restraints of 75 kcal mol^{−1} were imposed between KIR A and KIR B. Water molecules were added manually using $F_o - F_c$ electron density maps contoured at 3.0 σ . The final model includes residues 1–278 of HLA-Cw3, 0–99 of β 2m, 4–200 for both KIR molecules and all 9 residues of the peptide GAVDPLAL.

Structural analysis

Buried surface area calculations were performed with the program SURFACE⁴⁰ using a probe radius of 1.4 Å. Shape correlation statistics were determined using the program SC¹⁵. The hinge angle between various domains was determined using the program HINGE

(P.S., unpublished), which calculates the angle between vectors representing the principal moment of inertia for polyaniline models of each domain. The residues used for calculating the hinge angles are 4–103 and 107–200, respectively, for the KIR D1 and D2 domains, and 1–182 and 183–275, respectively, for the MHC $\alpha 1\alpha 2$ and $\alpha 3$ domains. The docking angle between KIR2DL2 and HLA-Cw3 was calculated using residues 4–200 of KIR2DL2 and 1–182 of HLA-Cw3.

SPR and analytical ultracentrifugation measurements

HLA-Cw3 was reconstituted in the presence of individual peptides as listed in Table 3. SPR measurements were performed using a BIAcore 2000 instrument (BIAcore AB). Either soluble KIR2DL2 receptor (0.1 mg ml⁻¹) or a class I MHC specific monoclonal antibody w6/32 (0.04 mg ml⁻¹) were immobilized on a CM5 sensor chip at pH 6.0 with *N*-hydroxysuccinimide / 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide hydrochloride (NHS/EDC). Immobilization of the receptor was assessed by the binding of the KIR2DL2 blocking antibody GL183 (Coulter Corp) and the integrity of refolded HLA-Cw3 was assessed by the binding of HLA-Cw3 to w6/32. The binding of HLA-Cw3 to immobilized KIR2DL2 was measured using serial dilutions of HLA from 0.3 to 40 μ M at a flow rate of 5 μ l min⁻¹ without regeneration of the chip. HLA-E was used as a negative control. The dissociation constants (K_d) were obtained from the steady-state curve fitting analysis using BIAevaluation 3.0 (BIAcore AB) or linear regression using ORIGIN 3.0 (MicroCal Software, Inc.).

Sedimentation equilibrium profiles were obtained in a Beckman XL-A at rotor speeds of 20K, 28K and 35K r.p.m. Samples with varying concentrations of HLA-Cw3 and KIR2DL2 in 20 mM sodium cacodylate and 20 mM CaCl₂ were used. All components were well described by thermodynamically ideal monomeric species, with buoyant molar masses consistent with those calculated from the amino-acid sequence. Data was analysed by a global fit of several equilibrium distributions at multiple rotor speeds, protein concentrations and molar ratios.

Generation of mutant KIR

A complementary DNA encoding residues 1–200 of KIR2DL2 was subcloned into pET30BirA allowing in-frame fusion of a sequence encoding the substrate for BirA. KIR mutations were then generated using a DNA QuikChange Kit (Stratagene) and plasmids containing the desired mutations were identified by automated sequencing.

Received 24 January; accepted 4 April 2000.

- Biron, C. A., Nguyen, K. B., Pien, G. C., Cousens, L. P. & Salazar-Mather, T. P. Natural killer cells in antiviral defense: function and regulation by innate cytokines. *Annu. Rev. Immunol.* **17**, 189–220 (1999).
- Lanier, L. L. NK cell receptors. *Annu. Rev. Immunol.* **16**, 359–393 (1998).
- Lanier, L. L., Corliss, B. C., Wu, J. J., Leong, C. & Phillips, J. H. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature* **391**, 703–707 (1998).
- Long, E. O. Regulation of immune responses through inhibitory receptors. *Annu. Rev. Immunol.* **17**, 875–904 (1999).
- Pende, D. *et al.* The susceptibility to natural killer cell-mediated lysis of HLA class I-positive melanomas reflects the expression of insufficient amounts of different HLA class I alleles. *Eur. J. Immunol.* **28**, 2384–2394 (1998).
- Fan, Q. R. *et al.* Structure of the inhibitory receptor for human natural killer cells resembles haematopoietic receptors. *Nature* **389**, 96–100 (1997).
- Snyder, G. A., Brooks, A. G. & Sun, P. D. Crystal structure of the HLA-Cw3 allotype-specific killer cell inhibitory receptor KIR2DL2. *Proc. Natl Acad. Sci. USA* **96**, 3864–3869 (1999).
- Maenaka, K., Juji, T., Stuart, D. I. & Jones, E. Y. Crystal structure of the human p58 killer cell inhibitory receptor (KIR2DL3) specific for HLA-Cw3-related MHC class I. *Structure* **7**, 391–398 (1999).
- Boyington, J. C. *et al.* Structure of CD94 reveals a novel C-type lectin fold: implications for the NK cell-associated CD94/NGK2 receptors. *Immunity* **10**, 75–82 (1999).
- Tormo, J., Natarajan, K., Margulies, D. H. & Mariuzza, R. A. Crystal structure of a lectin-like natural killer cell receptor bound to its MHC class I ligand. *Nature* **402**, 623–631 (1999).
- Winter, C. C., Gumperz, J. E., Parham, P., Long, E. O. & Wagtmann, N. Direct binding and functional transfer of NK cell inhibitory receptors reveal novel patterns of HLA-C allotype recognition. *J. Immunol.* **161**, 571–577 (1998).
- Winter, C. C. & Long, E. O. A single amino acid in the p58 killer cell inhibitory receptor controls the ability of natural killer cells to discriminate between the two groups of HLA-C allotypes. *J. Immunol.* **158**, 4026–4028 (1997).
- Biassoni, R. *et al.* Role of amino acid position 70 in the binding affinity of p50.1 and p58.1 receptors for HLA-Cw4 molecules. *Eur. J. Immunol.* **27**, 3095–3099 (1997).
- Fan, Q. R. & Wiley, D. C. Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor. *J. Exp. Med.* **190**, 1113–1123 (1999).
- Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950 (1993).
- Ysern, X., Li, H. & Mariuzza, R. A. Imperfect interfaces. *Nature Struct. Biol.* **5**, 412–414 (1998).
- Wang, J. H. *et al.* Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors. *Cell* **97**, 791–803 (1999).
- Mandelboim, O. *et al.* The binding site of NK receptors on HLA-C molecules. *Immunity* **6**, 341–350 (1997).
- Zappacosta, F., Borrego, F., Brooks, A. G., Parker, K. C. & Coligan, J. E. Peptides isolated from HLA-Cw*0304 confer different degrees of protection from natural killer cell-mediated lysis. *Proc. Natl Acad. Sci. USA* **94**, 6313–6318 (1997).
- Falk, K. *et al.* Allele-specific peptide ligand motifs of HLA-C molecules. *Proc. Natl Acad. Sci. USA* **90**, 12005–12009 (1993).
- Garboczi, D. N. *et al.* Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**, 134–141 (1996).
- Rajagopalan, S. & Long, E. O. The direct binding of a p58 killer cell inhibitory receptor to human histocompatibility leukocyte antigen (HLA)-Cw4 exhibits peptide selectivity. *J. Exp. Med.* **185**, 1523–1528 (1997).
- Vales-Gomez, M., Reyburn, H. T., Erskine, R. A. & Strominger, J. Differential binding to HLA-C of p50-activating and p58-inhibitory natural killer cell receptors. *Proc. Natl Acad. Sci. USA* **95**, 14326–14331 (1998).
- Maenaka, K. *et al.* Killer cell immunoglobulin receptors and T cell receptors bind peptide-major histocompatibility complex class I with distinct thermodynamic and kinetic properties. *J. Biol. Chem.* **274**, 28329–28334 (1999).
- Davis, S. J., Ikemizu, S., Wild, M. K. & van der Merwe, P. A. CD2 and the nature of protein interactions mediating cell-cell recognition. *Immunol. Rev.* **163**, 217–236 (1998).
- Colonna, M., Borsellino, G., Falco, M., Ferrara, G. B. & Strominger, J. L. HLA-C is the inhibitory ligand that determines dominant resistance to lysis by NK1- and NK2-specific natural killer cells. *Proc. Natl Acad. Sci. USA* **90**, 12000–12004 (1993).
- Mandelboim, O. *et al.* Protection from lysis by natural killer cells of group 1 and 2 specificity is mediated by residue 80 in human histocompatibility leukocyte antigen C alleles and also occurs with empty major histocompatibility complex molecules. *J. Exp. Med.* **184**, 913–922 (1996).
- Barber, L. D. *et al.* The inter-locus recombinant HLA-B*4601 has high selectivity in peptide binding and functions characteristic of HLA-C. *J. Exp. Med.* **184**, 735–740 (1996).
- Peruzzi, M., Parker, K. C., Long, E. O. & Malnati, M. S. Peptide sequence requirements for the recognition of HLA-B*2705 by specific natural killer cells. *J. Immunol.* **157**, 3350–3356 (1996).
- Rojo, S., Wagtmann, N. & Long, E. O. Binding of a soluble p70 killer cell inhibitory receptor to HLA-B*5101: requirement for all three p70 immunoglobulin domains. *Eur. J. Immunol.* **27**, 568–571 (1997).
- Littau, R. A. *et al.* An HLA-C-restricted CD8+ cytotoxic T-lymphocyte clone recognizes a highly conserved epitope on human immunodeficiency virus type 1 gag. *J. Virol.* **65**, 4051–4056 (1991).
- Hanke, T. *et al.* Direct assessment of MHC class I binding by seven Ly49 inhibitory NK cell receptors. *Immunity* **11**, 67–77 (1999).
- Orihuela, M., Margulies, D. H. & Yokoyama, W. M. The natural killer cell receptor Ly-49A recognizes a peptide-induced conformational determinant on its major histocompatibility complex class I ligand. *Proc. Natl Acad. Sci. USA* **93**, 11792–11797 (1996).
- Rajagopalan, S., Winter, C. C., Wagtmann, N. & Long, E. O. The Ig-related killer cell inhibitory receptor binds zinc and requires zinc for recognition of HLA-C on target cells. *J. Immunol.* **155**, 4143–4146 (1995).
- Davis, D. M. *et al.* The human natural killer cell immune synapse. *Proc. Natl Acad. Sci. USA* **96**, 15062–15067 (1999).
- Otwinski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
- Brunger, A. T. *et al.* Crystallography & NMR system: A Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in the models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- The Collaborative Computing Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1995).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Applied Crystallography* **24**, 946–950 (1991).
- Merritt, E. A. & Murphy, M. E. P. Raster 3D Version 2.0: a program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
- Nicholls, A., Chapp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).

Acknowledgements

We thank Z. Dauter and C. Titlow for help with synchrotron data collection; J. Lukasz for peptide synthesis; C. Hammer for mass spectroscopy measurements; M. Garfield for N-terminal amino-acid sequencing; A. Winterhalter and N. Mifsud for assistance with the mutagenesis and DNA sequencing; and D. Margulies for help with BIAcore SPR measurements. This work is supported by intramural research funding from the National Institute of Allergy and Infectious Diseases and an R. D. Wright Fellowship from the National Health and Medical Research Council, Australia.

Correspondence and requests for materials should be addressed to P.S. (e-mail: psun@nih.gov). Coordinates have been deposited in the Protein Data Bank under accession code 1EFX and are available before release from P.S.

Rossby waves on the Sun as revealed by solar 'hills'

J. R. Kuhn*, J. D. Armstrong*, R. I. Bush† & P. Scherrer†

* Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA

† Stanford University, HEPL 4085, Stanford, California 94305, USA

It is a long-standing puzzle that the Sun's photosphere—its visible surface—rotates differentially, with the equatorial regions rotating faster than the poles. It has been suggested¹ that waves analogous to terrestrial Rossby waves, and known as r-mode oscillations, could explain the Sun's differential rotation: Rossby waves are seen² in the oceans as large-scale (hundreds of kilometres) variations of sea-surface height (5-cm-high waves), which propagate slowly either east or west (they could take tens of years to cross the Pacific Ocean). Calculations show that the solar r-mode oscillations have properties that should be strongly constrained by differential rotation³. Here we report the detection of 100-m-high 'hills' in the photosphere, spaced uniformly over the Sun's surface with a spacing of $(8.7 \pm 0.6) \times 10^4$ km. If convection under the photosphere is organized by the r-modes, the observed corrugated photosphere is a probable surface manifestation of these solar oscillations.

The notion that global Rossby, or inertial, waves could exist within a rotating star was first described in calculations by Papaloizou and Pringle⁴. They named these toroidal oscillations r-modes, after Rossby waves. Unlike the intermediate period g-modes⁵, whose properties are determined by buoyancy restoring forces, the nearly incompressible r-modes are subject to Coriolis forces and have periods which are near the rotation period. They

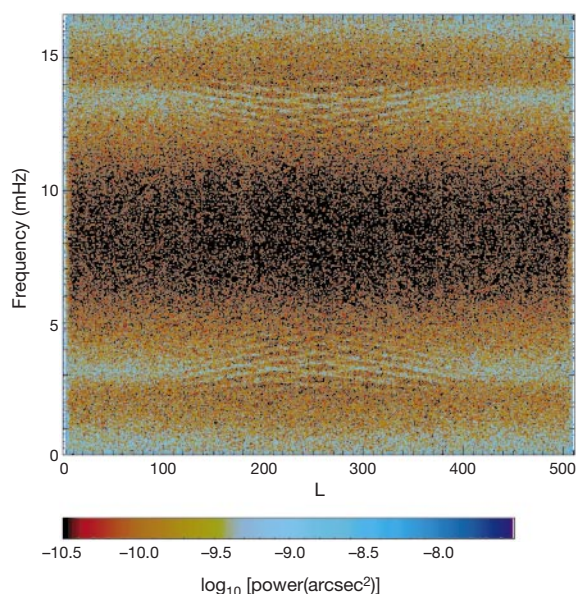


Figure 1 The two-dimensional power spectrum of the limb displacement. This was derived from an 8-h period of 1-min cadence data obtained in position angle bins 0.7° wide. The horizontal axis corresponds to the angular frequency (L , in units of waves per cycle) while the vertical axis plots the temporal frequency. We note that a measurement at angular frequency k tends to sample spherical harmonic modes of angular and azimuthal order equal to k ($l = m = k$). The image colourscale indicates the logarithm of the derived limb displacement power in units of arcsec^2 . This figure shows the complete spectrum, including Nyquist aliasing about the central temporal and spatial frequencies. The p-mode ridge structure near frequencies of 3 mHz is clearly observed here in the changing solar limb shape.

have not been convincingly observed, but detailed calculations^{3,6,7} imply that these modes are a potent probe of the largest structures in the solar convection zone, because they sample both the solar rotation and the large-scale convection timescales. With sufficient amplitude, r-modes will even affect a star's convection and differential rotation flows as suggested in ref. 1.

The Michelson Doppler Imager (MDI) on board the Solar and Heliospheric Observatory (SOHO) was designed primarily for Doppler observations, although it has produced unique astrometric data because of its accurate pointing and imaging capability from above the Earth's atmosphere. Recently, MDI improved on other attempts to measure the static solar limb shape by achieving an uncertainty of less than one milliarcsec (ref. 8) in the solar quadrupole and hexadecapole shape. Figure 1 illustrates the astrometric sensitivity of the technique—here the solar shape distortion produced by global p-modes has been detected. These modes (which

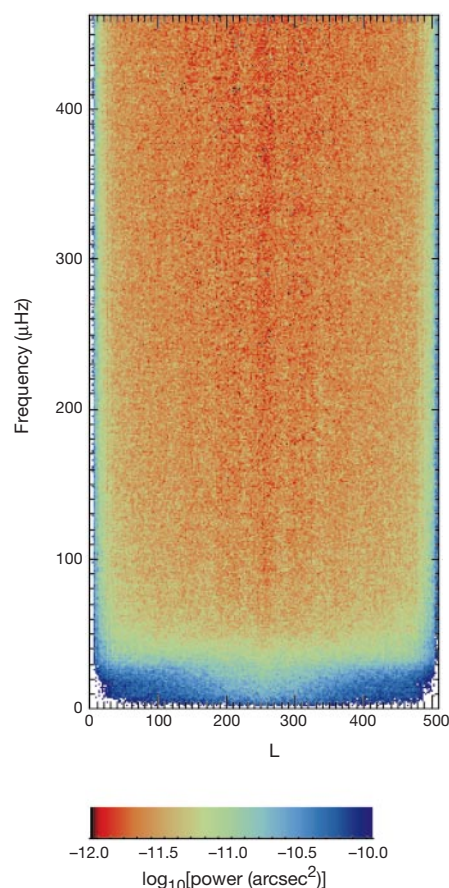


Figure 2 The two-dimensional, low-frequency, power spectrum of the MDI limb timeseries. Over a period of 26 months, a 6-pixel-wide annulus which includes the solar limb was generated every 12 min (ref. 9). These data were analysed as described in ref. 7 to reveal the limb position and brightness in 512 position angle bins covering the solar limb, and with 98,304 temporal samples at a cadence of 12 min. Gaps from telemetry problems and instrument mode changes corrupted $\sim 3\%$ of this timeseries. Measurements obtained after the SOHO disruption on 25 June 1998 and after the spacecraft reacquired the Sun on 16 September 1998 were also not included here because of a higher noise background caused by frequent spacecraft and experiment operating mode changes. The horizontal axis describes the angular frequency as obtained from the discrete transform with respect to the limb position angle. Slowly varying instrumental noise was minimized by removing a 180° wide boxcar average from each limb data record. An 8.3-d boxcar temporal average was also removed to yield a record of the residual temporal fluctuations in the solar limb over a 2.2-yr period. The two-dimensional Fourier transform yields the limb displacement power spectral density in units of arcsec^2 per frequency bin, which is displayed here on a logarithmic colourscale.

are of intermediate spherical harmonic degree, and which have 5-min periods) have radial amplitudes of about 15 cm s^{-1} so that one oscillation mode produces a limb displacement amplitude of about 10 microarcsec. This is consistent with the amplitude of the ridge structure displayed in Fig. 1 as it was measured from 8 hours of 1-min cadence MDI solar limb data.

Measurements of longer-period oscillations are possible using the nearly continuous MDI dataset obtained between 19 April 1996 and 24 June 1998 (ref. 10). Analysis of the MDI limb timeseries (Fig. 2) clearly shows excess low frequency power on all spatial scales below $25 \mu\text{Hz}$. The corresponding displacement velocity amplitude (Fig. 3) shows an obvious power excess near $20 \mu\text{Hz}$ with a root-mean-square velocity amplitude of about 0.3 mm s^{-1} . This is not likely to be caused by low radial and angular order solar g-modes because such modes more efficiently penetrate the convection zone to reach the photosphere at higher frequencies corresponding to periods of one to a few hours (ref. 5).

The temporal signature of a photospheric structure (for example, a standing wave pattern) as it traverses the limb due to solar rotation can be detected in the temporal power spectrum of the limb displacement, computed as a function of position angle (or solar latitude). The apparent sawtooth-shaped power enhancement in Fig. 4 is consistent with a co-rotating photospheric modulation having a latitude-independent transverse scale of $(8.7 \pm 0.6) \times 10^4 \text{ km}$. As expected for a solar (rather than an instrumental) phenomenon, the power distribution near the poles is noticeably smeared by the 7° inclination of the solar rotation axis from the normal direction to the ecliptic plane. We also analysed data from approximately one month when the MDI was rotated several degrees from its nominal roll angle. Data obtained during this period confirms that this signal is not associated with the MDI instrument, because the 'sawtooth' power distribution, calculated as in Fig. 4, rotated in apparent position angle by an amount in agreement with the new MDI roll angle.

This regular structure of 100-m-high photospheric 'hills', uniformly spaced over the surface of the Sun with a characteristic separation of approximately 90,000 km, is most plausibly explained as the co-rotating photospheric signature of long period r-modes. It is not surprising that these modes have been invisible from the

ground, as long-period oscillations are more readily detected from astrometric (shape) measurements from space than from even the most sensitive ground-based Doppler observations. At low frequencies, solar Doppler data are dominated by incoherent velocity noise from the convection zone. At low temporal frequencies, even a small velocity amplitude oscillation may be extracted from the convective noise in limb shape measurements because the limb displacement amplitude grows linearly with oscillation period (for a constant velocity amplitude), while the convection noise contributes only incoherently.

We now consider if this signal could be produced by the solar supergranulation¹¹. With a transverse scale of 30,000 km, the time-scale for limb perturbations due to these features would be 4 h, or about one-third of the period of the observed $20 \mu\text{Hz}$ power excess. We could propose that only the largest supergranule elements are detected at the limb because of foreshortening and projection effects. Then, depending on the statistical distribution of supergranules, we might find only one-in-two or one-in-three of them to be visible at the limb. With the assumption that supergranulation is a random convection process we expect the timeseries of limb perturbations to be random, but with a mean time between limb perturbations of, say, 11 h. This stochastic timeseries (called a "random telegraph signal"¹²) has a well-known power spectrum, $P(\omega)$, in the form of a lorentzian function, $P(\omega) = P_0/[1 + (\omega/\omega_0)^2]$, where P_0 and ω_0 are parameters that characterize the amplitude and mean frequency of the random telegraph signal. Thus, the existence of a mean timescale (11 h in this case) leads to a spectrum which is monotonic, and which shows a $1/\omega^2$ decline beyond the characteristic frequency. Only if there is long-range coherence in the

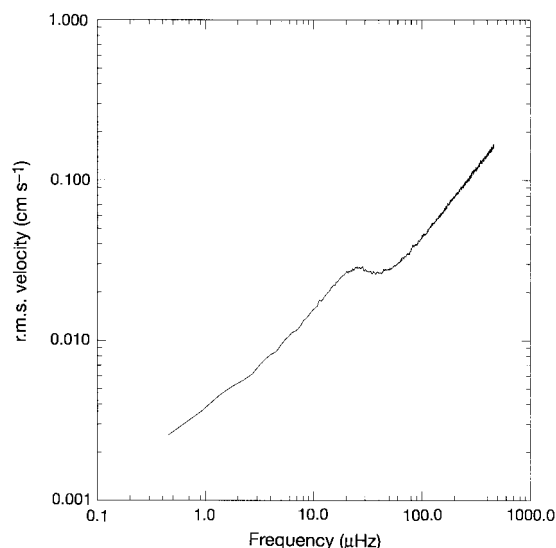


Figure 3 Root-mean-square limb shape velocity amplitude plotted against frequency. The average displacement power between angular frequencies of L and 256 has been converted to an r.m.s. velocity by scaling the amplitude by the temporal frequency. This is plotted on a logarithmic scale to reveal the mean effective velocity background from 800 d of astrometric observations. The excess power appears here as a bump at a frequency near $20 \mu\text{Hz}$ with a total r.m.s. amplitude of 0.03 cm s^{-1} .

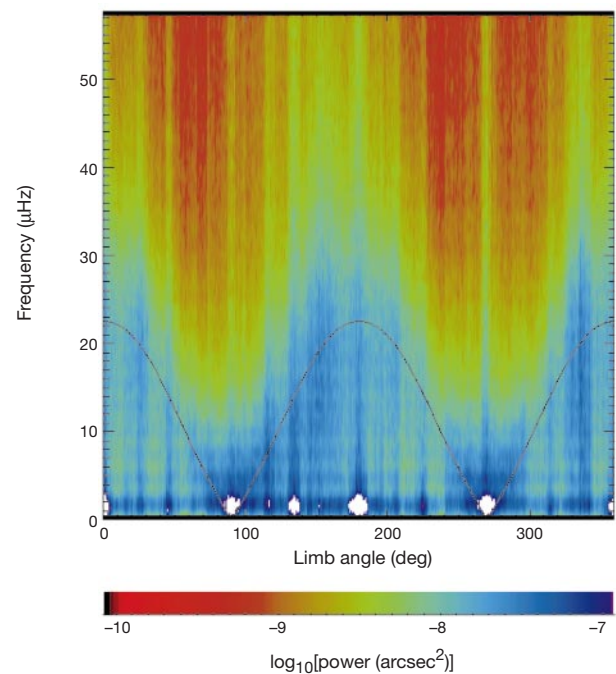


Figure 4 The power distribution plotted here shows the average power per temporal frequency bin versus position angle. The colourscale describes the average power in units of arcsec^2 per bin on a logarithmic intensity scale. The limb displacement was computed from the residual of a 256-bin running mean of each limb position record. Temporal residuals were then evaluated from the difference between this computed limb position and an 8.3-d running mean of the timeseries (duration, $7.1 \times 10^7 \text{ s}$). Position angle 0 corresponds to the west limb with increasing angles approaching the solar north pole. The power spectrum has been smoothed with an 88-point running boxcar average before plotting. The solid line shows the expected frequency of vertical photospheric modulation, with a transverse scale of $8.75 \times 10^4 \text{ km}$, which is rotating at the photospheric differential rotation rate.

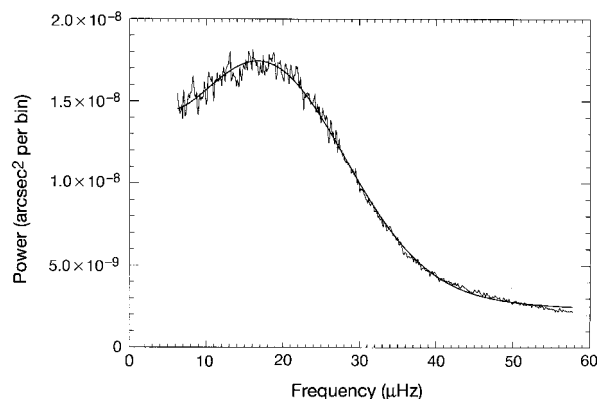


Figure 5 The mean power spectrum averaged over solar latitude. This was computed by assuming that the limb power spectra are due to a photospheric displacement pattern, rotating across the limb at various velocities corresponding to the local latitudinal rotation speed. The mean spectrum is obtained by first stretching the frequency scale of each latitudinal spectrum to match the equatorial spectrum (by linear interpolation onto the new frequency domain), and then averaging these on the common frequency domain. A simple shift of each spectrum by a frequency equal to the sawtooth variation with latitude (from Fig. 4) also matches the broad power excess at each latitude to the equatorial rate, yielding a mean spectrum which is essentially the same as the scaled frequency result displayed here. The solid line shows a least-squares fit to a gaussian and a power-law background distribution. The resulting centre frequency is 18 μHz .

timeseries (over many characteristic timescales) will the power spectrum yield a peak at the average temporal frequency. Thus, a test of the stochastic or supergranule hypothesis requires looking for a lorentzian form to the low-frequency spectrum.

The spectrum from a single latitude bin is too noisy to distinguish a monotonic excess power signal (a lorentzian) from the generally increasing displacement noise power towards lower frequencies. Combining spectra from a range of latitude bins improves the signal-to-noise ratio. We find a gaussian peak in the mean spectrum (Fig. 5) with a centre frequency of 18 μHz . The existence of a spectral peak suggests that at least a portion of the low-frequency power is not caused by supergranules. Although the spectrum also contains a displacement 'background' signal which increases to lower frequencies, the peak signal (corresponding to a 15-h period) is best described as a low- Q oscillator. It has long-range order and is not a stochastic 'random telegraph'. A standing wave is a natural mechanism for producing such a lattice of solar 'hills'. Wolff^{13,14} previously suggested that a large-scale photospheric nodal pattern generated by phase-locked r-modes would evolve over timescales of a few days—a result in qualitative agreement with the low- Q oscillator we see here in the mean limb displacement power spectrum. If r-modes are driven by the largest scales of convection, we should also expect a characteristic length scale of the order of the depth of the convection zone: that is, 2×10^5 km. The transverse scale we observe is half this, but is plausibly consistent.

The magnitude of the vertical displacement is also energetically reasonable. If subphotospheric convection is spatially organized by Rossby modes, then even velocities of 100 m s^{-1} (a small fraction of measured photospheric convective velocities) are energetic enough to lift up the overlying photosphere by the observed 100-m vertical height. Although it remains to be seen from detailed model calculations exactly how global solar r-modes could be excited, the existence of a corrugated photosphere with long-range order is good evidence for these oscillations. □

Received 16 December 1999; accepted 16 March 2000.

1. Plaskett, H. H. The polar rotation of the Sun. *Mon. Not. R. Astron. Soc.* **131**, 407–433 (1966).
2. Chelton, D. B. & Schlax, M. G. Global observations of oceanic Rossby waves. *Science* **272**, 234–238 (1996).

3. Wolff, C. L. Linear r-mode oscillations in a differentially rotating star. *Astrophys. J.* **502**, 961–967 (1998).
4. Papaloizou, J. & Pringle, J. E. Non-radial oscillations of rotating stars and their relevance to the short-period oscillations of cataclysmic variables. *Mon. Not. R. Astron. Soc.* **182**, 423–442 (1978).
5. Kumar, P., Quataert, E. & Bahcall, J. Observational searches for solar g-modes: some theoretical considerations. *Astrophys. J.* **458**, 83–L85 (1996).
6. Provost, J., Berthoméu, G. & Rocca, A. Low frequency oscillations of a slowly rotating star—quasi toroidal modes. *Astron. Astrophys.* **94**, 126–133 (1981).
7. Wolff, C. L. & Blizard, J. B. Properties of r-modes in the Sun. *Sol. Phys.* **105**, 1–15 (1986).
8. Kuhn, J. R., Bush, R., Scheick, X. & Scherrer, P. The Sun's shape and brightness. *Nature* **392**, 155–157 (1998).
9. Lazrek, M. *et al.* First results on p-modes from GOLF experiment. *Sol. Phys.* **175**, 227–246 (1997).
10. Scherrer, P. H. *et al.* The solar oscillations investigation—Michelson Doppler Imager. *Sol. Phys.* **162**, 129–188 (1995).
11. Schrijver, C. J., Hagenaar, H. R. & Title, A. M. On the patterns of the solar granulation and supergranulation. *Astrophys. J.* **475**, 328–337 (1997).
12. Rice, S. O. Mathematical analysis of random noise. *Bell Syst. Tech. J.* **23**, 1–162 (1944).
13. Wolff, C. L. Distinctive patterns on the surface of slowly rotating stars whose oscillations are non-linearly coupled. *Astrophys. J.* **193**, 721–727 (1974).
14. Wolff, C. Oscillation convection coupling: cause of supergranulation. *Astrophys. J.* **443**, 423–433 (1995).

Acknowledgements

We thank J. Saba, C. DeForest and J. Covington for assistance in operating the MDI instrument during these observations. We are particularly grateful to R. Bogart and to X. Scheick for their help with the limb software analysis.

Correspondence and requests for materials should be addressed to J.R.K. (e-mail: jkuhn@solar.stanford.edu).

Acceleration of quantum decay processes by frequent observations

A. G. Kofman & G. Kurizki

Department of Chemical Physics, The Weizmann Institute of Science, Rehovot 76100, Israel

In theory, the decay of any unstable quantum state can be inhibited by sufficiently frequent measurements—the quantum Zeno effect^{1–10}. Although this prediction has been tested only for transitions between two coupled, essentially stable states^{5–8}, the quantum Zeno effect is thought to be a general feature of quantum mechanics, applicable to radioactive³ or radiative decay processes^{6,9}. This generality arises from the assumption that, in principle, successive observations can be made at time intervals too short for the system to change appreciably^{1–4}. Here we show not only that the quantum Zeno effect is fundamentally unattainable in radiative or radioactive decay (because the required measurement rates would cause the system to disintegrate), but also that these processes may be accelerated by frequent measurements. We find that the modification of the decay process is determined by the energy spread incurred by the measurements (as a result of the time–energy uncertainty relation), and the distribution of states to which the decaying state is coupled. Whereas the inhibitory quantum Zeno effect may be feasible in a limited class of systems, the opposite effect—accelerated decay—appears to be much more ubiquitous.

We recall the common argument leading to the quantum Zeno effect (QZE)^{3,4}. If a system ruled by hamiltonian $\hat{H}$ is in state $|e\rangle$ at $t = 0$, the probability to find the system there at $t > 0$ is:

$$p_{ee}(t) = \left| \left\langle e \left| \exp \left(-\frac{i}{\hbar} \hat{H} t \right) \right| e \right\rangle \right|^2 \quad (1)$$

Provided the hamiltonian variance $\langle V^2 \rangle = \langle e | \hat{H}^2 | e \rangle - \langle e | \hat{H} | e \rangle^2$ is finite, then, at short times $\langle V^2 \rangle t^2 / \hbar^2 \ll 1$:

$$\rho_{ee}(t) \approx 1 - \langle V^2 \rangle t^2 / \hbar^2 \quad (2)$$

If instantaneous ideal measurements (projections) are performed at intervals τ , then $\rho_{ee}(t = n\tau) = \rho_{ee}^n(\tau)$. For a sufficiently large frequency of measurements $\nu \sim 1/\tau$, such that, at least, $\nu \gg \langle V^2 \rangle^{1/2} / \hbar$, it follows from equation (2) that:

$$\rho_{ee}(t = n\tau) \approx \exp[-(\langle V^2 \rangle \tau / \hbar^2) t] \quad (3)$$

Equation (3) may be limited to short times, such that $\rho_{ee}(t) \approx 1$, due to the possibility of transitions back from $|j\rangle$ to $|e\rangle$, which occurs, for example, when $|e\rangle$ is coupled to one^{5–8} or more discrete states. However, when $|e\rangle$ is weakly coupled to a broad band of states (continuum), the transitions back to $|e\rangle$ can be neglected and equation (3) has no time restriction. As follows from equation (3), the decay rate of state $|e\rangle$ decreases with τ (the QZE), the evolution becoming frozen in the limit $\tau \rightarrow 0$, which is known as the quantum Zeno paradox^{2–4}. The above argument seems compelling, and has led to the widespread opinion that the QZE is a necessary consequence of quantum mechanics, provided that the measurements performed on the system are ideal and sufficiently frequent (although a dissenting opinion has been voiced, based on a certain model of measuring the position of a tunnelling particle¹¹).

As shown below, the energy uncertainty (spread) incurred by frequent ideal projections fundamentally restricts the QZE observability, and renders decay acceleration by frequent measurements far more ubiquitous than its inhibition. We refer to this universal effect (first discovered for spontaneous emission in cavities¹²) as the anti-Zeno effect (AZE). By contrast, a fundamental limitation on the observability of the QZE emerges from our theory: the intervals τ between consecutive measurements required for the QZE are so short for many processes (such as radiative and radioactive decay) that the resulting energy spread may destroy the observed system by coupling it to unwarranted channels—in the worst case, leading to the creation of new particles (for example, electron–positron pair creation by a spontaneously emitting atom).

To get deeper insight into the effect of frequent measurements, it is necessary to go beyond the standard expansion, equation (2). We write $\hat{H} = \hat{H}_0 + \hat{V}$, so that $|e\rangle$ is an eigenvector of $\hat{H}_0$ with the eigenvalue $\hbar\omega_a = \langle e | \hat{H} | e \rangle$, whereas $\hat{V}$ describes the interaction of $|e\rangle$ with other states, $\hat{V} = \sum_j (V_{ej} |e\rangle\langle j| + V_{je} |j\rangle\langle e|)$, where $V_{ej} = \langle e | \hat{H} | j \rangle$ and $\{|j\rangle\}$ is an arbitrary orthonormal basis of the subspace of the Hilbert space orthogonal to $|e\rangle$ (formally, $\hat{V} = [\hat{P}, [\hat{P}, \hat{H}]]$, where $\hat{P} = |e\rangle\langle e|$). Choosing $|j\rangle$ to be the eigenvectors of $\hat{H}_0$ with the eigenvalues $\hbar\omega_j$, we can write the wavefunction of the system as:

$$|\Psi(t)\rangle = \alpha(t)e^{-i\omega_a t}|e\rangle + \sum_j \beta_j e^{-i\omega_j t}|j\rangle \quad (4)$$

One can then obtain from the Schrödinger equation the following equations

$$\dot{\alpha} = -\frac{i}{\hbar} \sum_j V_{ej} e^{-i(\omega_a - \omega_j)t} \beta_j \quad (5)$$

$$\dot{\beta}_j = -\frac{i}{\hbar} V_{je} e^{-i(\omega_a - \omega_j)t} \alpha \quad (6)$$

the initial condition being $|\Psi(0)\rangle = |e\rangle$. Formally integrating equation (6) yields:

$$\beta_j(t) = -\frac{i}{\hbar} V_{je} \int_0^t dt' e^{-i(\omega_a - \omega_j)t'} \alpha(t') \quad (7)$$

Inserting equation (7) into equation (5), we obtain the exact integro-differential equation

$$\dot{\alpha} = -\int_0^t dt' e^{i\omega_a(t-t')} \Phi(t-t') \alpha(t') \quad (8)$$

where

$$\Phi(t) = \hbar^{-2} \sum_j |V_{ej}|^2 e^{-i\omega_j t} = \hbar^{-2} \langle e | \hat{V} e^{-i\hat{H}_0 t / \hbar} \hat{V} | e \rangle \quad (9)$$

Equation (8) is exactly soluble by the Laplace transform method. However, for the present study of frequent measurements it suffices to obtain the short-time behaviour, when $\alpha(t) \approx \alpha(0) = 1$. This can be done iteratively. Setting $\alpha(t') = 1$ on the right-hand side of equation (8) yields (after the first iteration):

$$\alpha(t) \approx 1 - \int_0^t dt' (t-t') \Phi(t') e^{i\omega_a t'} \quad (10)$$

Note that as long as $\alpha(t) \approx 1$, powers of t higher than 2 can be important, even for $|V_{ej}|t \ll 1$, due to the exponent in equation (9), whence it follows that the standard expansion—equation (2)—may fail. Therefore the approximate result $\rho_{ee}(t) = |\alpha(t)|^2$, where $\alpha(t)$ is given by equation (10), is much more reliable than the commonly accepted^{3,4,9,10} equation (2), as it takes into account all powers of t .

If instantaneous projections onto $|e\rangle$ are performed at sufficiently small intervals τ , we can write, using equation (10)

$$\rho_{ee}(t = n\tau) = |\alpha(\tau)|^{2n} \approx \exp(-Rt) \quad (11)$$

where the measurement-modified decay rate is given by:

$$R = 2\text{Re} \int_0^\infty dt f(t) \Phi(t) \quad (12)$$

Here the measurement effects are accounted for by $f(t) = (1 - t/\tau) e^{i\omega_a t} \theta(\tau - t)$, where $\theta(x)$ is 1 for $x > 0$ and 0 for $x < 0$. The Fourier transform of $\Phi(t)$

$$G(\omega) = \frac{1}{\pi} \text{Re} \int_0^\infty dt \Phi(t) e^{i\omega t} = \hbar^{-2} \sum_j |V_{ej}|^2 \delta(\omega - \omega_j) \quad (13)$$

allows us to recast the expression for the measurement-modified decay rate (equation (12)) as

$$R = 2\pi \int_0^\infty d\omega G(\omega) F(\omega) \quad (14)$$

where

$$F(\omega) = \frac{\tau}{2\pi} \text{sinc}^2\left(\frac{(\omega - \omega_a)\tau}{2}\right) \quad (15)$$

The function $G(\omega)$ in equation (13) is the spectral density of the states $|j\rangle$ weighted with the coupling strengths $\hbar^{-2}|V_{ej}|^2$. If the states $|j\rangle$ belong to a spectrally dense band ('reservoir'), $G(\omega) \rightarrow |V(\omega)|^2 \rho(\omega)$, where $V(\omega)$ are the coupling constants to states $|\omega\rangle$, whose spectral density is given by $\rho(\omega)$. It is then called the reservoir coupling spectrum¹³.

Whereas $G(\omega)$ is a property of the hamiltonian $\hat{H}$, the form factor $F(\omega)$ (equation (15)) shows the broadening (width) of ω_a , incurred by frequent measurements, to be of the order of the measurement frequency $\nu \sim 1/\tau$. Thus, equation (14) represents a universal result: the decay rate of a frequently measured state $|e\rangle$ is simply the overlap of the reservoir coupling spectrum and the measurement-induced level width (Fig. 1, upper inset, Fig. 2a).

Equation (14) shows that the decay rate is determined by the form of $G(\omega)$ in a band of width ν around ω_a . This can be interpreted as resulting from the broadening of level $|e\rangle$, in accordance with the energy–time uncertainty relation:

$$\Delta E \Delta t \sim \hbar \quad (16)$$

Here ΔE is the energy uncertainty of $|e\rangle$, equal to the characteristic width of $F(\omega)$ times $\hbar$, and $\Delta t = 1/\nu$ is the average interval between

the measurements. The effect of instantaneous measurements (projections) is to dephase level $|e\rangle$ (refs 4, 14), that is, make the phase of state $|e\rangle$ completely random. This effect is analogous to collisional broadening¹³: phase randomization by collisions yields a linewidth equal to the collision frequency ν_0 , that is the same linewidth as in the case of an unstable state with lifetime $\Delta t = 1/\nu_0$.

A simple graphical analysis of equation (14) yields our three main conclusions.

(1) As shown in Fig. 1, upper inset, the QZE—that is, the reduction of the decay rate as the measurement (dephasing) rate ν increases—is generally obtained when the following inequalities are satisfied:

$$\nu \gg \Gamma_R, |\omega_a - \omega_M| \quad (17)$$

Here Γ_R is the reservoir width, and ω_M is the centre of gravity of $G(\omega)$. In this limit, we can make the approximation $G(\omega) \approx C\delta(\omega - \omega_M)$ in the integrand of R in equation (14), with $C = \int G(\omega)d\omega = \langle V^2 \rangle$. This approximation becomes exact in the case of resonant Rabi oscillations ($\Gamma_R = 0$, $\omega_a = \omega_M$), which explains why the QZE is observable in Cook's system for any ν (refs 5–8). More generally, this approximation holds for any $G(\omega)$ that falls off faster than $1/|\omega - \omega_M|$ on the wings. Then equation (14) yields the most general result for the QZE

$$R \approx 2C/\nu \quad (18)$$

where $1/\nu = \pi F(\omega_a)$. The flattening of the spectral peak of $G(\omega)$ by the broad function $F(\omega)$ in the convolution equation (14) (see Fig. 1, upper inset) is the origin of the QZE; that is, the reduction of R as compared to the measurement-free decay rate $\gamma_{\text{free}} = R(\nu \rightarrow 0) = 2\pi G(\omega_a)$, which is the 'golden rule' value¹³. To put it simply, if the system is probed frequently enough (as dictated by equation (17)), the QZE arises because the effective decay rate is averaged over all decay channels, many of which are weak, due to the energy uncertainty incurred by the measurements. The result, equation (18), is completely insensitive to the spectral shape of the reservoir: only its integral C matters (provided that equation (17) holds). The present analysis implies the possibility of inhibiting vibrational relaxation by the QZE in molecules and solids (Fig. 1).

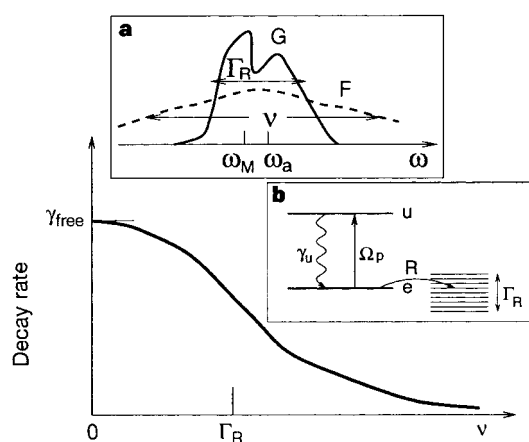


Figure 1 The QZE characteristics and a frequent-measurement scheme. Upper inset, conditions for the QZE (equation (17)) for a complicated reservoir-coupling spectrum $G(\omega)$ of width Γ_R (solid line) and measurement-induced level broadening $F(\omega)$ of width ν (dashed curve). Main figure, typical dependence of the decay rate R on ν under QZE conditions (equation (18)). Note that for decay near the cut-off frequency ω_c of a reservoir with $G(\omega) \propto (\omega_c - \omega)^{-\beta}$, where $0 < \beta < 1$ (for example, $\beta = 1/2$ for vibrational spectra with upper cut-off in molecules or solids), we find the QZE scaling to be slower, $R \propto \nu^{-\beta}$. Lower inset, scheme for frequent impulsive or continuous measurements of $|e\rangle$ (similar to refs 5, 6, 12).

(2) As shown in Fig. 2, left upper inset, the opposite behaviour is obtained in the limit

$$\nu \ll |\omega_m - \omega_a| \quad (19)$$

whenever ω_a is significantly detuned from the nearest maximum of $G(\omega)$ at ω_m , so that $G(\omega_a) \ll G(\omega_m)$. In this limit, the rate R grows with ν , because the dephasing function $F(\omega)$ is then probing more of the rising part of $G(\omega)$ in the convolution, equation (14). Hence, this limit should always yield a scaling of R as a positive power of ν : that is, the anti-Zeno effect (AZE) of decay acceleration by frequent measurements. Physically this means that, as the energy uncertainty increases with the measurement rate ν , the state decays into more and more channels, whose weight $G(\omega)$ is progressively larger.

The present analysis reveals the universality of the AZE: we may impose condition of equation (19) in any reservoir, as seen from Fig. 2, left upper inset. Moreover, the AZE is inevitable in natural decay processes, such as the nuclear β -decay: in such processes, any measurement rate ν below the relativistic cut-off of the reservoir, $\nu \ll Mc^2/\hbar$ (where Mc^2 is the rest-mass energy of the decaying particle), would correspond in equation (14) to probing the rising reservoir response $G(\omega) \propto \omega^\eta$ with $\eta \geq 1$. The result for $\eta > 1$ is then the following AZE dependence of the decay rate:

$$R \propto \nu \omega_R^{\eta-1} \quad (\omega_a/\omega_R^{\eta-1} \ll \nu \ll \omega_R) \quad (20)$$

Higher measurement rates, $\nu \gtrsim \omega_R$, are not only unfeasible, but also detrimental to the system, leading to the production of new particles.

Similar considerations apply to a two-level system, consisting of an upper state $|e\rangle$ at energy $\hbar\omega_e$ and a lower state $|g\rangle$ at energy $\hbar\omega_g$, with state $|e\rangle$ coupled to a zero-temperature reservoir of harmonic oscillators. This description fits atomic radiative decay (spontaneous emission) into a vacuum field^{13,15}. The effects of very frequent measurements, with $\nu \gtrsim \omega_a = \omega_e - \omega_g$, may be revealed in radiative decay of an excited hydrogenic (or Rydberg) level of a two-level

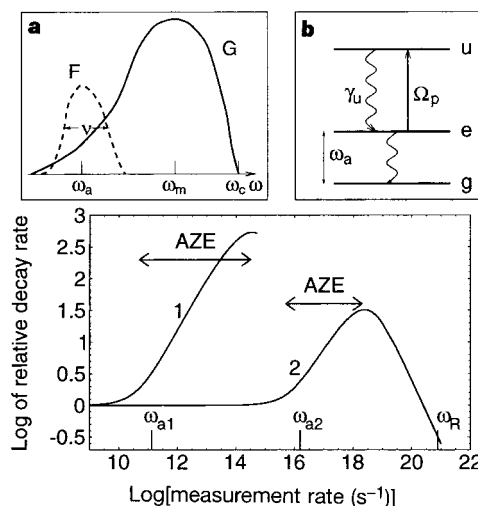


Figure 2 The AZE characteristics and a frequent-measurement scheme for two-level systems. **a**, Conditions for the AZE (equation (19)), with the same notation as in Fig. 1. Main figure, the dependence of the logarithm of the normalized decay rate $\log_{10}(R/\gamma_{\text{free}})$ on $\log_{10}\nu$ for a spontaneously emitting hydrogenic state. An exact calculation for this system¹⁵ provides the reservoir-coupling spectrum $G(\omega) = \alpha\omega/[1 + (\omega/\omega_b)^2]^4$, where α is determined by the field-atom coupling strength and $\omega_b \sim c/a_b$ is the non-relativistic cut-off frequency, a_b being the radius of the electron orbit. Under condition of equation (19) we then obtain from equations (14) and (15) the following AZE dependence: $R = \alpha\nu[\ln(\omega_b/\nu) + 0.354](\omega_a \ll \nu \ll \omega_b)$, where $\nu = 2/\tau$. The vertical lines on the horizontal axis denote the values of the atomic transition frequencies corresponding to curves 1, 2: $\omega_{a1} = 1.32 \times 10^{11} \text{ s}^{-1}$, $\omega_{a2} = 1.55 \times 10^{16} \text{ s}^{-1}$, whereas $\omega_R = 7.76 \times 10^{20} \text{ s}^{-1}$. The corresponding Bohr frequencies are $\omega_{b1} = 1.22 \times 10^{15} \text{ s}^{-1}$, $\omega_{b2} = 8.50 \times 10^{18} \text{ s}^{-1}$. The AZE ranges are marked. **b**, Scheme for realistic frequent measurements of the excited state in a two-level system^{5,6,12}.

system. As seen in Fig. 2, the AZE trend is revealed already for $\nu \lesssim \omega_a$, which should make its observation feasible, especially for microwave Rydberg transitions. A close inspection of Fig. 2 reveals that the decay rate R falls below the measurement-free (golden rule) decay rate γ_{free} only at $\nu \sim \omega_R$, which is not only unfeasible, but would imply a considerable probability of electron–positron pair creation. An additional limitation on the concept of measurement-influenced decay that is specific to two-level systems occurs for $\nu \gtrsim \omega_a$. The uncertainty of ω_a is then so large that occasionally $|g\rangle$ may be found at a higher energy than $|e\rangle$, which may cause $|g\rangle \rightarrow |e\rangle$ transitions. Such transitions are usually excluded by the rotating-wave approximation¹³, which may fail in this case because of the large energy fluctuations caused by the frequent measurements. Under these conditions, equation (11) holds only at short enough times, $t \ll R^{-1}$, for which $|g\rangle$ is still unpopulated.

(3) We can retrieve the limit in which the decay rate is unaffected by measurements and is given by the golden rule: $R \rightarrow \gamma_{\text{free}} = 2\pi G(\omega_a)$. In the case of a continuous $G(\omega)$, this limit is obtained from equation (14) if the measurement rate (dephasing) ν is too small (typically, $\nu \ll \omega_a$) to probe the slope of the spectral response. It is tantamount to the Wigner–Weisskopf approximation¹⁵ of purely markovian, exponential decay, $\rho_{ee}(t) \approx \exp(-\gamma_{\text{free}}t)$. This approximation is never strictly correct, because it requires an unphysical, spectrally flat reservoir $G(\omega) = \text{const.}$ ($-\infty < \omega < \infty$).

The arguments leading to equation (14) and the resulting effects have assumed ideal instantaneous projections on $|e\rangle$. Does a more realistic description of measurements still support these results? The answer is that their universality has been reaffirmed by detailed analysis of the two possible types of measurements of $|e\rangle$.

First, impulsive measurements^{5,6,12,16}. Such measurements are realizable, for example, when the decay process is repeatedly interrupted by a short electromagnetic pulse transferring the population of $|e\rangle$ to a higher auxiliary state $|u\rangle$, which then decays back to $|e\rangle$ incoherently (Fig. 1, lower inset, and Fig. 2b). For such a measurement to qualify as impulsive, its duration, which is of the order of the decay time of the $|u\rangle$ state $1/\gamma_u$ (assuming that the $|e\rangle \rightarrow |u\rangle$ transfer is faster than this decay), must be negligible, that is, it must be much shorter than the interval between successive measurements τ , which in turn should be much shorter than the decay time of $|e\rangle$. The solution of the full master equations for these measurements conforms to equation (14) to a very good approximation^{12,16}, with the dephasing function given by equation (15).

Second, continuous measurements. These should not be confused with continuous observation identified² as the limit of vanishing intervals between successive projections ($\Delta t \rightarrow 0$ or $\nu \rightarrow \infty$). This limit is unphysical, corresponding to an infinite spread of energy fluctuations. Even finite, but excessively large ν , may be detrimental to the model, as argued above. In contrast, realistic continuous measurements, through monitoring the state incessantly, still require a finite time for completing an observation; that is, they have a finite effective rate ν . This can be illustrated by the case of stationary dephasing when the $|e\rangle \leftrightarrow |u\rangle$ transition is driven by a near-resonant continuous-wave field at a Rabi frequency Ω that is much smaller than the transition linewidth $\gamma_u/2$ (Fig. 1, lower inset, and Fig. 2, upper right inset). It can be shown that equation (14) holds for continuous measurements with $F(\omega)$ being a lorentzian centred at ω_a of width $\nu = 2\Omega^2/\gamma_u$.

We may infer from the above discussion that, qualitatively, in all the cases considered above there is no essential distinction between different frequent measurements: ideal projections, impulsive measurements, and continuous measurements. We also note that, as the effect of both the impulsive and continuous measurements is tantamount to the dephasing of $|e\rangle$ at a rate ν , it is possible to mimic their effect by a stochastic (fluctuating) broadband field¹⁷. In all cases of realistic measurements, the spectral width ν of measurement-induced dephasing can be attributed to the time–energy uncertainty.

The simple criteria obtained from the universal formula, equation (14), lead us to conclude that the QZE—that is, decay inhibition by frequent measurements, whether ideal or more realistic—is strictly attainable only for decay into narrow reservoirs, whose spectral width is below the energy of the decaying state. This is the reason why the QZE is observable for Rabi oscillations^{5–8}, and is predicted here to be feasible for vibrational decay in molecules or solids (Fig. 2). Yet the QZE concept is untenable whenever the required broadening of the level by frequent measurements is so large that it has spectral overlap with transitions from the decaying state $|e\rangle$ to states outside the original reservoir. In particular, the QZE is incompatible with processes, such as the nuclear β -decay or spontaneous emission, wherein the reservoir response $G(\omega)$ grows with frequency almost up to the relativistic cut-off ω_R . The QZE requirement is then $\nu \gtrsim \omega_R$, and may result in the production of new particles (or other unwarranted effects specific to spontaneous emission, due to the breakdown of the rotating-wave approximation) and thus must be ruled out. By contrast, the condition of equation (19) for the accelerated decay (AZE) trend is in principle realizable for decay into any reservoir, whether spectrally broad or narrow, and is therefore far more ubiquitous than the QZE. This surprising conclusion can be experimentally tested: we predict that accelerated radiative decay is achievable for microwave transitions, $\omega_a \lesssim 10^{11} \text{ s}^{-1}$, in Rydberg atoms perturbed by randomizing collisions with a buffer gas at a rate $\nu \sim \omega_a$ or a fluctuating field with bandwidth $\nu_f \gtrsim \omega_a$ (Fig. 2). Another prediction is AZE for β -decay, according to equations (19) and (20), which may be achievable by perturbing the decaying nuclear state with a broadband (fluctuating) γ -ray source. Thus the present findings may serve as clues to our ability to manipulate diverse decay processes.

The essential restrictions on the QZE observability found here evoke one of Zeno's paradoxes (the 'Achilles') that has not been discussed in the context of quantum mechanics. This paradox arises if we assume that time is divisible *ad infinitum*. In this paradox, the running Achilles can never catch the crawling tortoise in front of him, because he must first reach where the tortoise started. But when he reaches there, the tortoise has moved ahead, and so on. Quantum measurement theory supports Zeno's reservations regarding infinitesimal time intervals, because they exact an impossible toll: energy spread tending to infinity (according to the time–energy uncertainty relation), with detrimental consequences to the system below a certain minimal time interval. □

Received 26 August 1999; accepted 8 March 2000.

1. Khalifin, L. S. Phenomenological theory of K^0 mesons and the non-exponential character of the decay. *JETP Lett.* **8**, 65–68 (1968).
2. Misra, B. & Sudarshan, E. C. G. The Zeno's paradox in quantum theory. *J. Math. Phys.* **18**, 756–763 (1977).
3. Sakurai, J. J. *Modern Quantum Mechanics* 484–486 (Addison-Wesley, Reading, Massachusetts, 1994).
4. Joos, E. Continuous measurement: Watchdog effect versus golden rule. *Phys. Rev. A* **29**, 1626–1633 (1984).
5. Cook, R. J. What are quantum jumps? *Phys. Scr. T* **21**, 49–51 (1988).
6. Itano, W. M., Heinzen, D. J., Bollinger, J. J. & Wineland, D. J. Quantum Zeno effect. *Phys. Rev. A* **41**, 2295–2300 (1990).
7. Knight, P. L. Watching a laser hot-pot. *Nature* **344**, 493–494 (1990).
8. Frerichs, V. & Schenzle, A. Quantum Zeno effect without collapse of wave packet. *Phys. Rev. A* **44**, 1962–1968 (1991).
9. Schulman, L. S. Continuous and pulsed observations in the quantum Zeno effect. *Phys. Rev. A* **57**, 1509–1515 (1998).
10. Facchi, P. & Pascazio, S. Temporal behavior and quantum Zeno time of an excited state of the hydrogen atom. *Phys. Lett. A* **241**, 139–144 (1998).
11. Fearn, H. & Lamb, W. E. Jr Computational approach to the quantum Zeno effect: Position measurements. *Phys. Rev. A* **46**, 1199–1205 (1992).
12. Kofman, A. G. & Kurizki, G. Quantum Zeno effect on atomic excitation decay in resonators. *Phys. Rev. A* **54**, R3750–R3753 (1996).
13. Cohen-Tannoudji, C., Dupont-Roc, J. & Grynberg, G. *Atom-Photon Interactions* (Wiley, New York, 1992).
14. Milburn, G. J. Quantum Zeno effect and motional narrowing in a two-level system. *J. Opt. Soc. Am. B* **5**, 1317–1322 (1988).
15. Moses, H. E. Exact electromagnetic matrix elements and exact selection rules for hydrogenic atoms. *Lett. Nuovo Cimento* **4**, 51–53 (1972).
16. Beige, A. & Hegerfeldt, G. C. Quantum Zeno effect and light-dark periods for a single atom. *J. Phys. A* **40**, 1323–1334 (1997).

17. Harel, G., Kofman, A. G., Kozhekin, A. & Kurizki, G. Control of non-Markovian decay and decoherence by measurements and interference. *Opt. Express* **2**, 355–367 (1998).

Acknowledgements

G.K. is the G. Dunne Professor of Chemical Physics, A.K. is supported by the Israel Ministry of Absorption. This work was supported by the ISF, Minerva and EU (TMR).

Correspondence and requests for materials should be addressed to G.K. (e-mail: Gershon.Kurizki@weizmann.ac.il).

Absence of thermodynamic phase transition in a model glass former

Ludger Santen & Werner Krauth

CNRS-Laboratoire de Physique Statistique, Ecole Normale Supérieure,
24 rue Lhomond, 75231 Paris Cedex 05, France

The glass transition can be viewed simply as the point at which the viscosity of a structurally disordered liquid reaches a universal threshold value¹. But this is an operational definition that circumvents fundamental issues, such as whether the glass transition is a purely dynamical phenomenon². If so, ergodicity gets broken (the system becomes confined to some part of its phase space), but the thermodynamic properties of the liquid remain unchanged across the transition, provided they are determined as thermodynamic equilibrium averages over the whole phase space. The opposite view^{3–6} claims that an underlying thermodynamic phase transition is responsible for the pronounced slow-down in the dynamics at the liquid–glass boundary. Such a phase transition

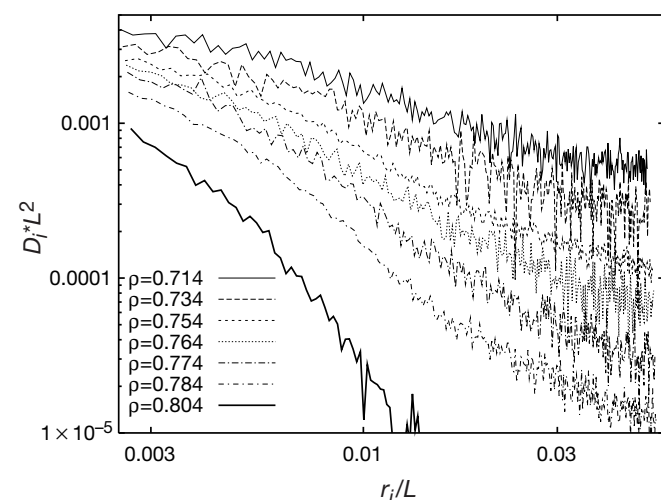


Figure 1 Effective diffusion constants as a function of the radius for polydisperse hard disks. The upper 6 curves represent 256 disks with polydispersity $\Delta/r_1 = 19$ simulated using the local-move Monte Carlo algorithm at densities for which the algorithm equilibrated. The effective diffusion constants were calculated using $D_i = \langle (\mathbf{x}_i(t_0 + t) - \mathbf{x}_i(t_0))^2 \rangle / (4t)$, for $|\mathbf{x}_i(t_0 + t) - \mathbf{x}_i(t_0)| \gg r_i$. The time was measured in Monte Carlo sweeps. Only the motion relative to the largest disk was taken into account. These asymptotic constants agree (up to a rescaling of the time unit) with what would be obtained in a molecular dynamics simulation⁸. The lowest curve is as above, but equilibrated initial configurations were provided by the cluster Monte Carlo algorithm. We observe that the smallest 50 disks are able to pass through the blocked matrix made up of the largest 200 disks. These findings are consistent with the estimated values of the transition densities. L , side of box containing disk. r_i , radius of disk i .

would trigger the dynamic standstill, and then be masked by it. Here we perform Monte Carlo simulations of a two-dimensional system of polydisperse hard disks far within its glassy phase. The approach⁷ allows for non-local moves in a way that preserves micro-reversibility. We find no evidence for a thermodynamic phase transition up to very high densities; the glass is thus indistinguishable from the liquid on purely thermodynamic grounds.

We considered polydisperse hard disks $i = 1, \dots, N$ with radii r_i (where $r_i - r_{i-1} = \Delta/(N-1)$) in a square box of volume $V = L^2$ with periodic boundary conditions. The polydispersity Δ/r_1 was kept fixed and the system was studied as a function of the density $\rho = \pi \sum_i r_i^2 / L^2$. We note that hard-core systems are athermal: the phase diagram and the dynamics (up to a trivial rescaling) are independent of temperature. The external control parameters are the density (for the [NV] ensemble) or the pressure (for [NP]).

Conventional, local-move Monte Carlo simulations in the [NV] ensemble were used to compute effective diffusion constants⁸ at densities for which the local algorithm was still ergodic (Fig. 1). We failed to detect any finite-size effects by comparing runs with 256 and 1024 disks at density $\rho = 0.764$, but we noticed strong dependence of the diffusivities on the disk size. For each disk i we found its diffusivity D_i to agree very well with⁹

$$D_i(\rho) \approx (\rho - \rho_i^G)^\alpha \quad (1)$$

For the 180 largest disks we found no systematic dependence of $\rho_i^G = \rho^G \approx 0.805 \pm 0.01$ (with a best fit $\alpha \approx 2.4$). For the smallest disks i , the extrapolated values for ρ_i^G increased slowly. Our result for ρ^G agrees with what was found in a related system¹⁰.

We next performed simulations with the cluster Monte Carlo algorithm⁷. There, groups of disks are swapped around a ‘pivot’ in a way which represents an alternative Markov-chain sampling of the Boltzmann distribution. The correctness of both implementations was validated by comparing structural quantities (pair correlation functions) and thermodynamic variables in the manifestly liquid regime, where the local algorithm still converges well. We found that the cluster Monte Carlo algorithm does not slow down as we pass ρ^G .

We used the cluster Monte Carlo algorithm to compute the equation of state in the [NP] ensemble⁸. For pressures corresponding to the glass transition density ρ^G , we reached a relative precision for V of 0.05% during a one-week simulation on a single processor workstation ($N = 256$). The compressibility of the system was obtained both by deriving the algebraic fit to $V(P)$ (see Fig. 2

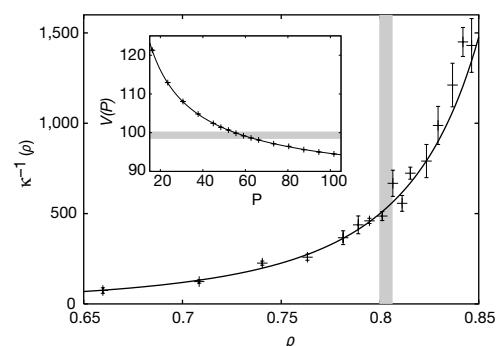


Figure 2 Equilibrium compressibility. The system, as before, was simulated using the cluster Monte Carlo algorithm. The equilibrium equation of state (inset) and inverse compressibility κ^{-1} (main figure) were obtained in the [NP] ensemble. κ was obtained from the fluctuations of the volume (points with error bars) and from the derivative of the best $V(P)$ fit according to the functional form $V(P) - k = aP^{-b}$ (k is a constant; the parameters of the best fit are $k = 84.3$, $a = 254$, $b = 0.69$). Indications of an equilibrium phase transition are lacking. The grey lines indicate the localization of the glass transition as extrapolated from Fig. 1.

inset) and by computing the volume fluctuations:

$$\kappa = \frac{-1}{\langle V \rangle} \frac{\partial \langle V \rangle}{\partial P} = \frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V \rangle^2} \quad (2)$$

The agreement of the two methods (Fig. 2) testifies to the excellent convergence of the cluster Monte Carlo algorithm. Some of the compressibility data at high density were obtained from several independent runs with identical results. In contrast, the local Monte Carlo algorithm yielded non-reproducible (generally lower) compressibilities.

The results presented in Fig. 2 allow us to draw the following conclusions:

- (1) In the polydisperse hard-disk system considered, the glass is indistinguishable from the liquid on purely thermodynamic grounds, as we find no indication of a phase transition.
- (2) Furthermore, the glass remains thermodynamically stable. No tendency towards crystallization or phase separation was detected. At smaller values of the polydispersity Δ/r_1 , a crystalline state is present, and can be readily detected in the equation of state¹¹. In a related two-dimensional system of binary mixtures, phase separation was evidenced by the cluster Monte Carlo algorithm¹².
- (3) Finally, the smooth behaviour of the compressibility as we cross ρ^G indicates that the ergodicity of the cluster Monte Carlo algorithm is preserved up to the highest densities studied.

Beyond conclusion (3) above, the cluster Monte Carlo algorithm has passed an extremely stringent ergodicity test in a system with 15 polydisperse disks (where we also identified a density ρ^G of complete

dynamical standstill). A large number of equilibrated configurations were stored during single cluster simulation runs in the [NV] ensemble at various densities ρ . For each stored configuration we then alternated very small up-scalings of each radius $r_i \rightarrow (1 + \epsilon)r_i$ ($i = 1, \dots, N$) with a few local Monte Carlo moves ('rattling'). The procedure was repeated until a jamming state (with unchanged polydispersity Δ/r_1) was reached. This state represents an inherent structure (IS), generalized from what is done in thermal systems by a quench to zero temperature^{13,14}. Histograms of the IS densities ρ_{IS} for different values of ρ show that the algorithm explores the remaining regions of phase space as the density ρ is increased (Fig. 3).

A non-ergodic algorithm would get stuck in one or a few IS (during a single run) as regions of phase space become dynamically inaccessible. We observed the contrary: almost identical sets of IS with extremely large values of $\rho_{IS} \gg \rho$ appeared in simulations at different, very high values of ρ (for intermediate values of ρ_{IS} , the match could no longer be achieved, simply because of the very large number of different IS, even in a system with 15 disks (Fig. 3)). Figure 4a shows a configuration at $\rho = 0.86$, which can be blown up into the densest IS which we were able to find at polydispersity $\Delta/r_1 = 19$ ($\rho_{IS} = 0.8938$). This IS (like many others) recurred frequently during a single run both at $\rho = 0.84$ and at $\rho = 0.82$. We conjecture that the IS related to Fig. 4a solves the closest-packing problem for $N = 15$ and $\Delta/r_1 = 19$. The configurations shown in Fig. 4a and Fig. 4b are connected by a move of the cluster Monte Carlo algorithm.

Our results show that the new Monte Carlo algorithm (which can be generalized to incorporate a smooth potential¹⁵) works extremely well at densities at which the local method is totally stuck. Besides the calculation of thermodynamic quantities, the algorithm can also be used to study the finite-time dynamics (starting from equilibrated samples) at $\rho \geq \rho^G$. This is how the lowest curve in Fig. 1 was obtained. As the initial condition is thermalized, we gain access to unbiased dynamical information in a regime which used to be out of reach of rigorous simulations because not all disks can be properly equilibrated with the local algorithm. The method represents a very powerful tool for the study of glasses. □

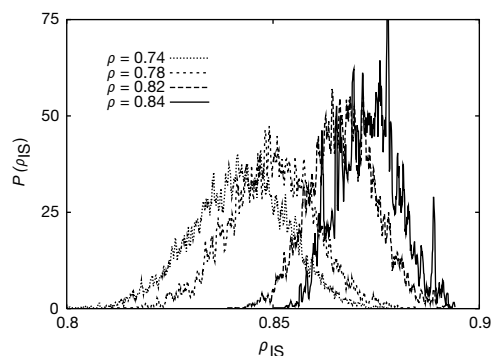


Figure 3 Distribution of inherent structure densities for 15 disks with polydispersity $\Delta/r_1 = 19$. Equilibrium configurations were generated during a long run of the cluster Monte Carlo algorithm. The corresponding inherent structures were obtained by iterating a rescaling of the particle radii and local Monte Carlo moves ('rattling') until a jamming state (of density ρ_{IS}) was reached. As ρ increases, inherent structures with higher ρ_{IS} become more probable. The same inherent structures with very large values of ρ_{IS} were found in the simulation at $\rho = 0.82$ and $\rho = 0.84$.

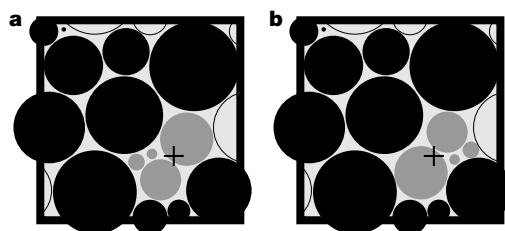


Figure 4 Example of a cluster move at high density. The crosses denote the randomly chosen 'pivot' with respect to which the cluster (grey disks) can be flipped⁷ from **a** to **b**. Periodic boundary conditions are indicated. The inherent structure related to the configuration in **a** is conjectured to yield the closest packed state for $N = 15$, $\Delta/r_1 = 19$.

Received 16 December 1999; accepted 16 March 2000.

1. Angell, C. A. Formation of glasses from liquids and biopolymers. *Science* **267**, 1924–1935 (1995).
2. Götze, W. & Sjögren, L. Relaxation processes in supercooled liquids. *Rep. Prog. Phys.* **55**, 241–376 (1992).
3. Kauzmann, W. The nature of the glassy state and the behavior of liquids at low temperatures. *Chem. Rev.* **43**, 219–256 (1948).
4. Gibbs, J. H. & DiMarzio, E. A. Nature of the glass transition and the glassy state. *J. Chem. Phys.* **28**, 373–383 (1958).
5. Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
6. Menon, N. & Nagel, S. R. Evidence for a divergent susceptibility at the glass transition. *Phys. Rev. Lett.* **74**, 1230–1233 (1995).
7. Dress, C. & Krauth, W. Cluster algorithm for hard spheres and related systems. *J. Phys. A* **28**, L597–L601 (1995).
8. Hansen, J. P. & McDonald, I. R. *Theory of Simple Liquids* (Academic, London, 1986).
9. Fuchs, M., Götze, W. & Mayr, M. R. Asymptotic laws for tagged-particle motion in glassy systems. *Phys. Rev. E* **58**, 3384–3399 (1998).
10. Doliwa, B. Dynamik harter Kugeln am Glasübergang. Diploma thesis, Univ. Mainz (1999).
11. Alder, B. J. & Wainwright, T. E. Phase transition in elastic disks. *Phys. Rev.* **127**, 359–361 (1962).
12. Buhot, A. & Krauth, W. Phase separation in two-dimensional additive mixtures. *Phys. Rev. E* **59**, 2939–2941 (1999).
13. Stillinger, F. H. & Weber, T. A. Packing structures and transitions in liquids and solids. *Science* **225**, 983–989 (1984).
14. Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
15. Malherbe, J. G. & Amokrane, S. Asymmetric mixture of hard particles with Yukawa attraction between unlike ones: a cluster algorithm simulation study. *Mol. Phys.* **97**, 677–683 (1999).

Acknowledgements

We thank C. Alba-Simionesco, J.-L. Barrat and A. Heuer for very useful discussions. L.S. acknowledges support from the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to W.K. (e-mail: krauth@physique.ens.fr).

Dynamics of the silicon (111) surface phase transition

J. B. Hannon*, H. Hibino†, N. C. Bartelt‡, B. S. Swartzentruber§, T. Ogino† & G. L. Kellogg§

* Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

† NTT Basic Research Laboratories, Atsugi, Kanagawa 243-0198, Japan

‡ Sandia National Laboratories, Livermore, California 94551, USA

§ Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

The manner in which phase transformations occur in solids determines important structural and physical properties of many materials. The main problem in characterizing the kinetic processes that occur during phase transformations is the difficulty of observing directly, in real time, the growth of one phase at the expense of another. Here we use low-energy electron microscopy to study the real-time kinetics of a phase transformation confined to the silicon (111) surface. We show that the transformation is governed by the rate at which material is exchanged between the first layer of the crystal and the surface. In bulk phase transformations, the dynamics are usually governed either by the rate of diffusion of material to the phase boundaries or by the structural rearrangement of atoms at the phase boundary¹. The kinetic process that we have identified here has no bulk analogue and leads to domain dynamics that are qualitatively different from those expected for bulk systems.

The phase transition we have investigated is the transformation between two crystallographic arrangements on a Si(111) surface, namely from the 7×7 to the 1×1 structure. Surface phase transitions involve changes in the arrangement of atoms at the outermost layers of a solid, but not the atoms within the bulk of the material. We are able to follow the surface transformation *in situ* using low-energy electron microscopy (LEEM)². At temperatures above 1,125 K, the surface undergoes a first-order^{3,4} phase transition from the reconstructed 7×7 structure to an unreconstructed 1×1 phase^{5,6}. The low-temperature phase is a complicated reconstruction with a large unit cell⁷. The 1×1 structure consists of a dense,

disordered structure on top of the bulk-terminated surface⁸. The structural transformation in this system is qualitatively similar to bulk precipitation: the density of the 1×1 phase is 6% larger than that of the low-temperature 7×7 phase⁹. Conversion of the 7×7 structure to the 1×1 structure requires the incorporation of additional Si at the boundaries between the phases.

A central question is where this material comes from. Reflection electron microscopy studies suggest that surface steps, which are observed to retract during the transition¹⁰, are the ultimate source of the additional silicon required for the conversion. This suggests a direct transfer of material from the retracting step to the boundary of the surface structure, or domain (Fig. 1a). We find, however, that there is an important intermediate process that completely determines how the domains evolve in time. Specifically, we find evidence that the random creation of atom–vacancy pairs on the surface provides the additional Si required for the transition. The atoms diffuse to the domain boundaries, and the vacancies diffuse to the step edges, leading to the observed retraction of surface steps (Fig. 1b). Our measurements show the evolution of these two domains during the transformation, from which we infer that domain growth that is qualitatively different from that observed in bulk systems. The essential difference is that the source of material required for the phase transition to proceed is continuously supplied to the surface from the underlying surface layer.

A LEEM image from the Si(111) surface 10 K above the phase transition temperature is shown in Fig. 2. The image was formed from 10-eV electrons scattered specularly from the surface. The reflectivity of the 7×7 regions is higher than that of the 1×1 regions when 10-eV electrons are used. Consequently, the 7×7 regions appear bright in the LEEM images. The initial domain configuration was formed by rapidly cooling the Si(111) surface below the transition temperature, after brief heating to 1,500 K to clean the surface. Upon cooling the 7×7 domains nucleate first at the upper side of step edges. Isolated domains have domain boundaries in the $\langle \bar{1}12 \rangle$ directions¹¹, and are usually triangular. As well as isolated 7×7 domains and 7×7 domains at step edges, domains of 1×1 structure completely surrounded by 7×7 regions are also observed (Fig. 2). Above the phase transition temperature, the structure shown in

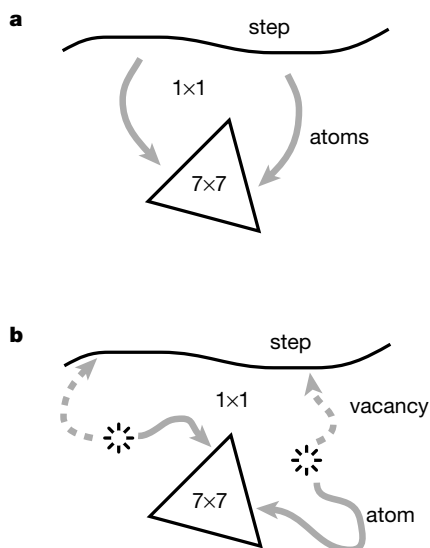


Figure 1 Processes leading to 7×7 domain decay and step retraction. **a**, atoms evaporate from step edges and incorporate at phase boundaries. **b**, Random atom–vacancy creation on the surface produces atoms that diffuse to phase boundaries, and vacancies that annihilate at step edges.

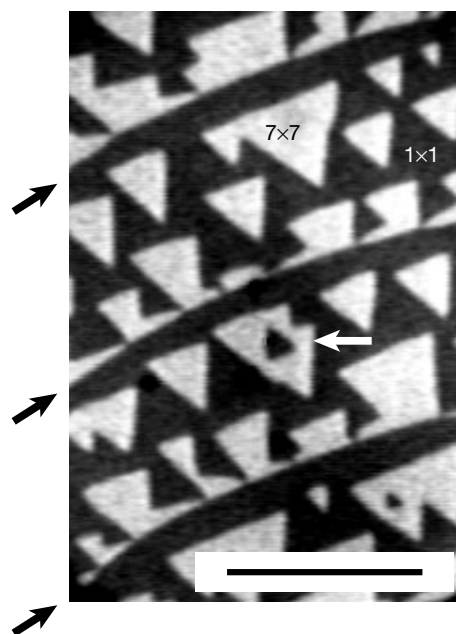


Figure 2 Bright-field 10 eV LEEM image of the Si(111) surface at 1,130 K with 7×7 and 1×1 domains. Dark arrows indicate step positions. A 1×1 domain completely surrounded by the 7×7 structure is indicated by a white arrow. Scale bar, 1 μm .

Fig. 2 is not stable. The fraction of the surface covered by the 7×7 structure decreases with time as the conversion to the 1×1 structure proceeds. The evolution of the domain structure is illustrated in the series of images shown in Fig. 3.

By measuring the time evolution of the domain sizes, the kinetics of the domain decay can be probed. In each frame of a sequence of images, the domain areas are determined by fitting polygons to the domain boundaries. The time evolution of the areas of selected domains in Fig. 3 is shown in Fig. 4. The data show marked variation in the decay rate among the domains. The variation suggests that the incorporation rate of atoms at the domain boundary is fast compared to surface diffusion. If the incorporation rate were small, gradients in the concentration of atoms on the surface would not develop. Every domain would be surrounded by the same atom density, and would decay at the same rate.

The most notable feature of the data in Fig. 4 is the fact that the decay rate is independent of the domain size. The decay rate of an individual domain is essentially constant throughout the decay of the domain. Surprisingly, the decay rate is strongly correlated with the initial size of the domain. Domains that are large when the decay begins have larger decay rates throughout the decay. As is shown below, the decay rate is determined by the local environment of the domain: the locations of neighbouring domains and bounding step edges.

All of these features of the data are reproduced in detail by a mathematical model that assumes the thermal generation of atom–vacancy pairs at the surface. In the limit where the atom–vacancy creation rate is small compared to the hop rate of atoms and vacancies at the surface, the process can be modelled as a constant flux of atoms and vacancies onto the surface. The decay of isolated 7×7 domains is therefore analogous to the growth of two-dimensional islands from a flux^{12,13}. The diffusion equation governing the concentration of atoms at the surface, c_a , is given by:

$$D_a \nabla^2 c_a + R_1 - R_2 c_a c_v = 0 \quad (1)$$

where R_1 is the rate at which atoms and vacancies are created, and R_2 is the probability that an atom and vacancy at the surface recombine. An identical equation governs the surface vacancy concentration, c_v . In this model the only source of Si is the atom–vacancy creation process. The quantities R_1 and R_2 define the equilibrium

concentrations of atoms and vacancies at the surface, $\bar{c}_a$ and $\bar{c}_v$. In terms of these quantities, equation (1) becomes:

$$D_a \nabla^2 c_a + R_1 \left(1 - \frac{c_a c_v}{\bar{c}_a \bar{c}_v} \right) = 0 \quad (2)$$

In the limit $c_a c_v \ll \bar{c}_a \bar{c}_v$, equation (2) reduces to the diffusion equation with a constant flux term equal to R_1 . Physically, this limit is reached when the atomic hop rate at the surface is fast compared to the atom–vacancy creation rate.

We have used this model to analyse the domain decay shown in Fig. 4. We assume that atom diffusion is sufficiently fast that the atom–vacancy creation process can be represented by a constant flux, F , incident on the 1×1 regions of the surface. From measurements of the evolution of 1×1 domains completely surrounded by 7×7 areas (described below) we conclude that atom–vacancy creation on 7×7 regions of the surface can be neglected.

The decay rate of a domain is equal to the rate at which atoms reach the phase boundary from the surrounding terrace:

$$\frac{dA}{dt} = \omega \oint D_a \nabla c_a \cdot \hat{n} dl \quad (3)$$

where A is the area of the domain, ω is the area of an atom, and D_a is the surface diffusion constant for atoms. The integration is carried out over the perimeter of the island. This model assumes that the domain boundaries are perfect sinks for the diffusing atoms. The absence of Ostwald ripening in the evolution of the domains, where large domains grow at the expense of small domains, supports this assumption. The atom concentration on the terrace is determined by solving the diffusion equation for atoms:

$$D_a \nabla^2 c_a + F = 0 \quad (4)$$

where F is a uniform, time-independent flux of atoms onto the surface. We applied the boundary condition where the atom concentration vanishes at domain boundaries.

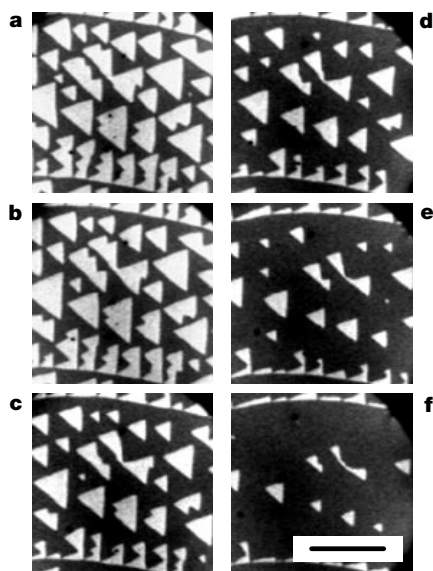


Figure 3 Sequence of bright-field 10-eV LEEM images of the Si(111) surface at 1,135 K. The time interval between images is 50 s, starting at panel **a**. Step edges run horizontally near the top and bottom of each frame. Scale bar, 1 μm .

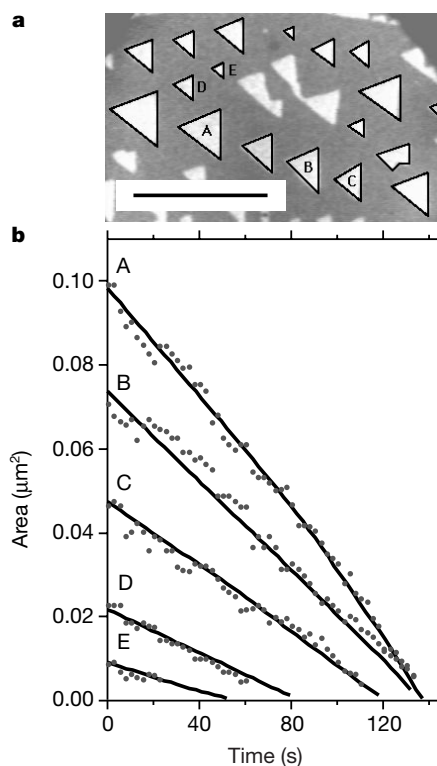


Figure 4 Decay rates for selected 7×7 domains. **a**, Marked domains from Fig. 3d. Scale bar, 1.0 μm . **b**, Filled circles indicate measured domain areas. Solid lines are the simulated decay curves using the model described in the text with $f = 1.0 \times 10^{-4} \text{ ML s}^{-1}$.

We make the further simplifying assumption that 7×7 domains at the step edge are static, but are still sinks for diffusing atoms. Static domains make the model easier to implement, and are supported by the experimental fact that domains at step edges decay much more slowly than those on the terrace (Fig. 3). It is clear from equation (4) that the concentration on the terrace is proportional to F/D_a . Consequently, the decay rate of a domain (equation (4)) will be equal to a characteristic rate ωF multiplied by a geometric factor independent of D_a . The only free parameter in this model is the flux, which determines the overall time scale of the decay. The relative decay rates of the domains are determined solely by the initial configuration and cannot be adjusted. In the simulations, equation (4) is solved for the initial configuration of domains. The decay rate of each domain is determined from equation (3). The domain sizes are adjusted on the basis of the computed decay rates. The diffusion equation is then solved for the new configuration, and the process is repeated to generate the area-versus-time curve for each domain. During the decay, the shape of each domain in the simulation is fixed. A comparison between the measured and simulated domain decay is shown in Fig. 4. The best agreement between simulation and experiment corresponds to $\omega F = 1.04 \times 10^{-4}$ monolayers (ML) per second. One monolayer is defined as the density of atoms at the bulk-terminated (111) surface. Assuming reasonable values for the prefactor, this rate corresponds to an activation energy for atom–vacancy generation in the range 3.2 to 3.8 eV. The model clearly reproduces both the constant decay rate of individual domains as well as the observed variation in the decay rates among the domains. The value of the flux can also be estimated from the data in Fig. 4. Conversion of the initial coverage of 7×7 domains, roughly 0.25 ML, requires that 0.015 ML of additional Si be supplied to the surface. The conversion occurs in about 100 s, yielding an effective flux of about 1.5×10^{-4} ML s^{-1} , in good agreement with the simulation.

The observed decay cannot be reproduced in models where steps are the primary source of additional Si. In particular, the correlation of the domain decay rate with initial domain size is not reproduced.

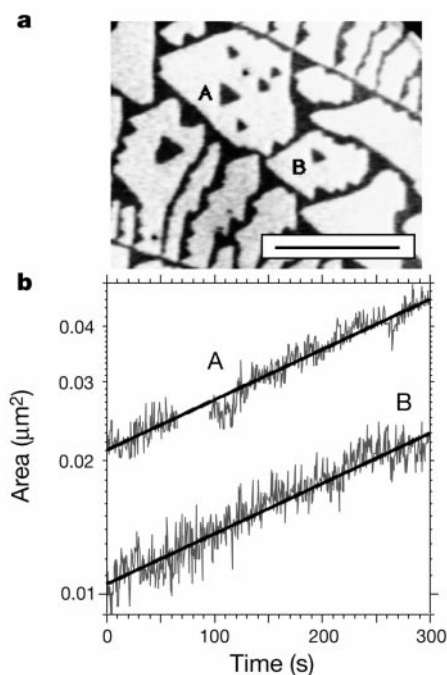


Figure 5 Growth of 1×1 domains above the phase transition temperature. **a**, Bright-field 10-eV LEEM image of 1×1 domains surrounded by 7×7 regions at 1,140 K. Scale bar, $1.0 \mu\text{m}$. **b**, Areas of the indicated domains as a function of time. Light lines are the measured areas. Heavy lines are the simulated areas with $F = 1.6 \times 10^{-4}$ ML s^{-1} .

In these models domains with direct line-of-sight to the steps decay the most quickly.

A central assumption of the successful model is that there is no atom–vacancy creation on the 7×7 regions of the surface. This assumption can be tested by measuring the growth of 1×1 regions surrounded by 7×7 regions (Fig. 5). Because diffusing atoms do not cross the domain boundaries, the model predicts that the growth rate of a ‘trapped’ 1×1 domain is determined by the flux of atoms into the domain itself:

$$\frac{dA}{dt} \approx \omega FA \quad (5)$$

suggesting that these should grow exponentially with time. Furthermore, because each trapped 1×1 domain is effectively isolated, a common growth rate should describe all domains, in sharp contrast to the variation in decay rates observed for the 7×7 domains. The measured decay rates of two trapped 1×1 domains are shown in Fig. 5b. For this particular sample temperature, analysis of the simultaneous decay of a 7×7 domain yields a flux of $\omega F = 1.6 \times 10^{-4}$ ML s^{-1} . The solid lines in Fig. 5b show the computed area of the trapped domains using the same value of the flux. The agreement shows that the atom–vacancy model describes the time evolution of both types of domains with a unique flux.

In our model, both vacancies and atoms are produced at the surface. If the vacancy concentration is not significantly supersaturated, domain evolution will be governed by the undersaturated atom concentration. In order for the vacancy concentration to remain close to equilibrium, a sink for vacancies at the surface is required. We propose that surface steps act as sinks for the vacancies. Vacancy annihilation at step edges leads to the step retraction observed by Latyshev *et al.* (ref. 10).

On the basis of simple bond counting, a lower atom–vacancy creation energy at surfaces compared to the bulk should be common, so that a process that is usually neglected in the bulk (such as atom–vacancy pair creation) can strongly influence surface mass transport. As the present work shows, reduced barriers at surfaces can lead to morphological evolution that is qualitatively different from that of the bulk. □

Received 12 November 1999; accepted 15 March 2000.

- Christian, J. W. in *Physical Metallurgy* (ed. Cahn, R. W.) 443–539 (North-Holland, Amsterdam, 1970).
- Bauer, E. Low energy electron microscopy. *Rep. Prog. Phys.* **57**, 895–938 (1994).
- Osakabe, N., Yagi, K. & Honjo, G. Reflection electron microscope observations of dislocations and surface structure phase transition on clean (111) silicon surfaces. *J. J. Appl. Phys.* **19**, L309–L312 (1980).
- Teliép, W. & Bauer, E. The $(7 \times 7) \leftrightarrow (1 \times 1)$ phase transition on Si(111). *Surf. Sci.* **162**, 163–168 (1985).
- Lander, J. J. Chemisorption and ordered surface structures. *Surf. Sci.* **1**, 125–164 (1964).
- Bennett, P. A. & Webb, M. W. The Si(111) 7×7 to 1×1 transition. *Surf. Sci.* **104**, 74–104 (1981).
- Takayanagi, K., Tanishiro, Y., Takahashi, S. & Takahashi, M. Structure analysis of the Si(111)- 7×7 reconstructed surface by transmission electron microscopy. *Surf. Sci.* **164**, 367–392 (1985).
- Becker, R. S., Golovchenko, J. A., Higashi, G. S. & Swartzentruber, B. S. New reconstructions on silicon (111) surfaces. *Phys. Rev. Lett.* **57**, 1020–1023 (1986).
- Yang, Y. N. & Williams, E. D. High atom density in the 1×1 phase and origin of the metastable reconstructions on Si(111). *Phys. Rev. Lett.* **72**, 1862–1865 (1994).
- Latyshev, A. V., Krasilnikov, A. B., Aseev, A. L., Sokolov, L. V. & Stenin, S. I. Reflection electron microscopy study of the clean Si(111) surface reconstruction during the $(7 \times 7) \leftrightarrow (1 \times 1)$ phase transition. *Surf. Sci.* **254**, 90–96 (1991).
- Yagi, K. & Tanishiro, Y. Studies of surface phase transition by electron microscopy. *J. Phase Trans.* **53**, 197–204 (1995).
- Bartelt, M. C., Schmid, A. K., Evans, J. W. & Hwang, R. Q. Island size and environment dependence of adatom capture: Cu/Co islands on Ru(0001). *Phys. Rev. Lett.* **81**, 1901–1904 (1998).
- Hannon, J. B., Bartelt, M. C., Bartelt, N. C. & Kellogg, G. L. Etching of the Si(001) surface with molecular oxygen. *Phys. Rev. Lett.* **81**, 4676–4679 (1998).

Acknowledgements

We thank W. W. Mullins and R. S. Sekerka for discussions. The experiments were performed at Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the US Department of Energy.

Correspondence and requests for materials should be addressed to J.B.H. (e-mail: jhannon@andrew.cmu.edu).

Old radiocarbon ages in the southwest Pacific Ocean during the last glacial period and deglaciation

Elisabeth L. Sikes^{*†}, Catherine R. Samson^{†‡}, Thomas P. Guilderson^{§||} & William R. Howard[†]

^{*} School of Environmental and Marine Sciences and Department of Geology, University of Auckland Private Bag 92019 Auckland, New Zealand

[†] Antarctic Cooperative Research Centre, GPO Box 252-80, Hobart, Tasmania, Australia

[‡] Institute for Antarctic and Southern Ocean Studies, University of Tasmania, GPO Box 252-77, Hobart, Tasmania, Australia

[§] Centre for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory L-397, 7000 East Avenue, Livermore, California 94551, USA

^{||} Department of Earth & Planetary Sciences, 20 Oxford Street, Harvard University, Cambridge, Massachusetts 02138, USA

Marine radiocarbon (^{14}C) dates are widely used for dating oceanic events and as tracers of ocean circulation, essential components for understanding ocean–climate interactions. Past ocean ventilation rates have been determined by the difference between radiocarbon ages of deep-water and surface-water reservoirs, but the apparent age of surface waters (currently ~ 400 years in the tropics and $\sim 1,200$ years in Antarctic waters¹) might not be constant through time², as has been assumed in radiocarbon chronologies^{3,4} and palaeoclimate studies⁵. Here we present independent estimates of surface-water and deep-water reservoir ages in the New Zealand region since the last glacial period, using volcanic ejecta (tephras) deposited in both marine and terrestrial sediments as stratigraphic markers. Compared to present-day values, surface-reservoir ages from 11,900 ^{14}C years ago were twice as large (800 years) and during glacial times were five times as large (2,000 years), contradicting the assumption of constant surface age. Furthermore, the ages of glacial deep-water reservoirs were much older (3,000–5,000 years). The increase in surface-to-deep water age differences in the glacial Southern Ocean suggests that there was decreased ocean ventilation during this period.

Surface ocean reservoir ages reflect a balance between equilibration with the atmosphere, radioactive decay, and the input of deep waters to the mixed layer of the ocean⁶. Whereas subtropical surface waters remain at the surface long enough to be in steady state with atmosphere⁶, thermohaline circulation removes surface water in the North Atlantic, isolating it from contact with the atmosphere. This causes its ^{14}C age to increase until the water outcrops in the Southern Ocean. There it partially re-equilibrates with the atmosphere before leaving the surface again; it forms Antarctic deep-water masses flowing northward, with initial radiocarbon ages of ~ 700 years for intermediate waters and $\sim 1,400$ years for bottom waters⁷. Subsurface return flow from the north in the Pacific causes the oldest reservoir ages of $\sim 1,500$ –2,000 years to occur at mid-depths⁷. Past changes in atmospheric ^{14}C levels that are not caused by changes in cosmogenic production are assumed to be owing to changes in ocean–atmosphere partitioning; these may be caused by changes in vertical mixing, ventilation of thermocline and deep waters, and changes in residence times related to changes in thermohaline circulation associated with colder, glacial-type climate conditions.

Unlike benthic-planktonic ^{14}C difference estimates of apparent ventilation ages^{5,8}, our approach to determining past reservoir ages uses tephras as stratigraphic marker beds to reference surface- and deep-water ^{14}C ages directly to the atmosphere using terrestrial ^{14}C (ref. 2). Volcanic eruptions on the North Island of New Zealand have produced tephras deposited on land with datable organic

matter incorporated within or bracketing the tephras⁹ and in adjacent ocean basins associated with planktonic and benthic foraminifera^{9,10} (Fig. 1). Marine ages were determined for this study by accelerator mass spectrometry radiocarbon (AMS ^{14}C) dating of foraminifera directly above and below the ashes in marine cores from the Bay of Plenty and the Chatham Rise (Fig. 1 and Table 1). The tephras in this study form layers 1–7 cm thick of nearly pure ash grains and contain features which indicate minimal bioturbation: sharp basal contacts below layers of nearly 100% ash, fining upwards of grains, and/or shower-bedding features characteristic of airfall deposition¹¹. The existence of pure tephra layers above a sharp contact indicates that the ash falls capped the sediments and although the sediments above and below the layers are clearly bioturbated, these layers remained separated by the ash. Differences between ^{14}C dates above and below the tephras confirm minimal mixing through the ash layers (Table 1). We use previously established mean terrestrial ^{14}C dates as reference ages for the tephras⁹. The surface-reservoir ages are the difference between planktonic foraminiferal and terrestrial ^{14}C ages, and the deep-water ages (or apparent ventilation ages) are the differences between benthic foraminiferal and terrestrial ages.

During the Holocene (Whakatane, 4,830 ^{14}C yr BP; Mamaku, 7,250 ^{14}C yr BP; and Rotoma, 8,530 ^{14}C yr BP tephras), surface-reservoir ages in the subtropical Bay of Plenty are indistinguishable from pre-bomb values of 400–470 years (ref. 12; Tables 2 and 3; Fig. 2). At 4,830 ^{14}C yr BP the apparent ventilation age at 1,675 m was 1,520 years. This is similar to modern deep-water radiocarbon

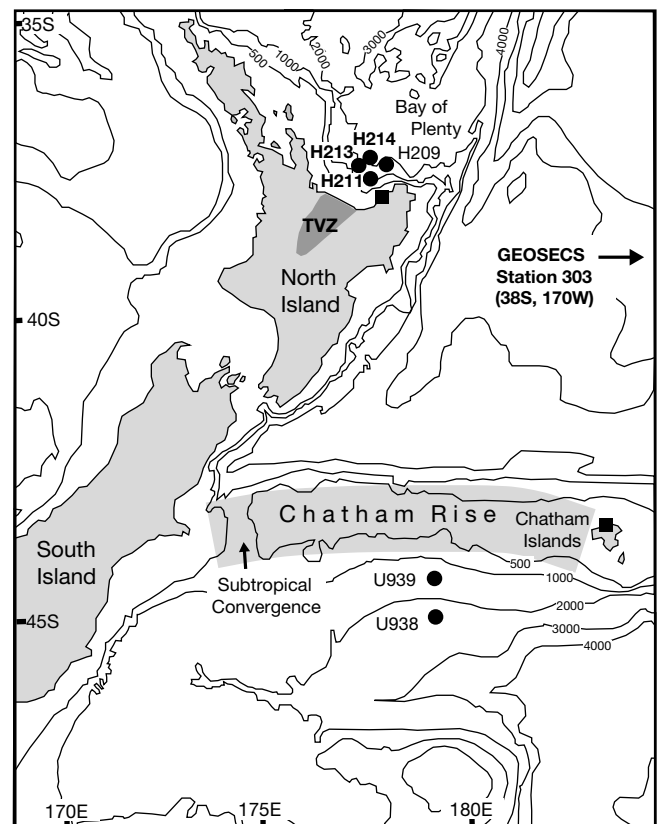


Figure 1 Locations of sediment cores and modern samples used in this study. Deep-sea sediment cores are from the Bay of Plenty and south of Chatham Rise (solid circles), and modern pre-bomb samples (solid squares) are coastal samples from the northeast North Island and the Chatham Islands. We note that the subtropical convergence (STC), which separates subtropical waters in the north from subpolar waters to the south, lies north of the southern Chatham Rise core sites. It is well established that on the Chatham Rise, the STC did not move from this position in the last glaciation³¹, in contrast to other parts of the Southern Ocean. The Taupo volcanic zone (TVZ) is the source of the tephras. Latitude of GEOSECS Station 303 (radiocarbon profile shown in Fig. 3) is indicated by arrow.

Table 1 Ash layer dates

Depth (cm)	Tephra layer	Sample	Lab code	¹⁴ C age	¹⁴ C error
H209 37° 09.5' S 177° 44.4' E 1,675 m					
57–58	Whakatane (4,830 ± 20)	<i>Gl. inflata</i>	CAMS 4172	4,960	50
57–58		<i>Gl. inflata</i>	CAMS 41743	4,880	40
58–60					
60–61		<i>Gl. inflata</i>	CAMS 41744	5,520	50
60–61		<i>Gl. inflata</i>	CAMS 41745	5,700	40
79–80	Rotoma (8,530 ± 10)*	<i>Gl. inflata</i>	CAMS 47187	13,280	60
80–81		<i>Gl. inflata</i>	CAMS 41746	13,160	50
81–82					
171–172	Waiohau (11,850 ± 60)	<i>Gl. inflata</i>	CAMS 40462	12,730	50
171–172		<i>Gl. inflata</i>	CAMS 40461	12,650	70
171–172		Mixed benthics	CAMS 41747	13,710	60
172–176					
176–177		<i>Gl. inflata</i>	CAMS 40463	12,640	50
176–177		<i>Gl. inflata</i>	CAMS 40464	12,720	50
176–177		Mixed benthics	CAMS 41748	13,280	50
H211 37° 18.2' S 177° 21.5' E 1,500 m					
88–91	Waiohau (11,850 ± 60)	<i>Gl. inflata</i>	CAMS 39603	12,130	50
90–95					
95–96		<i>Gl. inflata</i>	NZA 6668	12,750	160
95–96		<i>Gl. inflata</i>	CAMS 39604	12,650	70
95–96		<i>Gl. inflata</i>	CAMS 39605	12,620	50
158–161	Rerewhakaaitu (14,700 ± 110)	<i>Gl. inflata</i>	CAMS 41749	14,800	50
158–161		<i>Gl. inflata</i>	CAMS 41750	14,720	50
159–161		<i>Gl. inflata</i>	NZA 6665	14,740	130
167–170					
172–173		<i>Gl. inflata</i>	NZA 6666	15,330	150
H213 37° 02.5' S 177° 10.5' E 2,065 m					
60–61	Whakatane (4,830 ± 20)	<i>Gl. inflata</i>	CAMS 41751	3,650†	40
61–62					
62–63		<i>Gl. inflata</i>	CAMS 41752	5,120	50
62–63		<i>Gl. inflata</i>	CAMS 41753	5,110	40
62–63		Mixed benthics	CAMS 52010	6,350	80
240–243	Kawakawa (22,590 ± 230)	Mixed benthics	CAMS 41779	26,500	160
241–242		<i>Gl. inflata</i>	CAMS 41756	24,240	130
241–242		<i>Gl. inflata</i>	CAMS 41757	24,530	120
242–243		<i>Gl. inflata</i>	CAMS 41754	24,370	150
243–248					
248–249		<i>Gl. inflata</i>	CAMS 41755	24,770	150
248–251		Mixed benthics	CAMS 41778	25,610	140
H214 36° 55.5' S 177° 26.5' E 2,045 m					
75–77	Mamaku (7,250 ± 20)	<i>Gl. inflata</i>	NZA 6654	7,640	90
76–77		<i>Gl. inflata</i>	CAMS 39597	7,540	40
77–78					
78–79		<i>Gl. inflata</i>	CAMS 39598	7,610	40
78–79		<i>Gl. inflata</i>	CAMS 39599	7,700	40
94–96	Rotoma (8,530 ± 10)	<i>Gl. inflata</i>	CAMS 39600	8,800	30
96–97					
97–98		<i>Gl. inflata</i>	CAMS 39602	8,960	40
97–98	Waiohau (11,850 ± 60)	<i>Gl. inflata</i>	CAMS 39601	9,130	40
151–153		<i>Gl. inflata</i>	NZA 6655	12,820	110
153–159					
159–160		<i>Gl. inflata</i>	NZA 6662	12,910	140
160–161		<i>Gl. inflata</i>	CAMS 40465	12,940	70
202–203	Rerewhakaaitu (14,700 ± 110)	<i>Gl. inflata</i>	CAMS 40466	14,980	70
203–205		<i>Gl. inflata</i>	NZA 6663	14,980	120
205–206					
206–208		<i>Gl. inflata</i>	NZA 6664	15,350	160
208–209		<i>Gl. inflata</i>	CAMS 40467	14,970	70
208–209		<i>Gl. inflata</i>	CAMS 40468	14,910	70
U938 45° 04.5' S 179° 29.9' E 2,700 m					
127–131	Kawakawa (22,590 ± 230)				
131–133		<i>Gl. inflata</i>	CAMS 41778	24,390	120
130–134		<i>G. bulloides</i>	CAMS 50998	23,880	220
130–134		Mixed benthics	CAMS 50996	27,410	220
130–134		Mixed benthics	CAMS 50997	27,820	200
U939 44° 29.7' S 179° 30' E 1,300 m					
76–83	Kawakawa (22,590 ± 230)				
84.5–85		<i>Gl. inflata</i> 110 µg	CAMS 40472	24,810	510
84.5–85		<i>G. bulloides</i> 190 µg	CAMS 40473	24,870	310
84.5–85		<i>Uvigerina</i> 170 µg	CAMS 40474	25,590	380

Planktonic and benthic foraminiferal AMS ¹⁴C results. Samples were taken directly above and below tephra in the Bay of Plenty and south of the Chatham Rise. Planktonic foraminiferal ¹⁴C ages were generated on monospecific samples of *Globobulimina inflata* and in one sample on *Globigerina bulloides* (U939 84.5–85 cm; CAMS 40473). Both species are known to calcify in surface waters³⁹. Benthic foraminiferal ¹⁴C ages are based on mixed species. All samples were taken within 3 cm of the tephra, except above the Rerewhakaaitu in H211. Dates are conventional ¹⁴C ages based on a half life of 5,568 ± 30 years (ref. 1). Analyses were made at the Rafter Radiocarbon Laboratory, New Zealand (NZA) and the Centre for Accelerator Mass Spectrometry (CAMS), Lawrence Livermore National Laboratory, California, USA. Sub-optimal target weights are given in micrograms of carbon where necessary. Duplicates generally agree within the reported errors. There is no significant difference between the dates from Rafter Radiocarbon Laboratory and those from Lawrence Livermore National Laboratory, although Rafter errors are larger. Tephra identifications were based on relative position in the core, identification of the ferromagnesian assemblage and the chemical variations within the titanomagnetites¹¹, with the exception of the tephra at 90–95 cm in core H211 and 77–78 cm in core H214 for which the ¹⁴C between cores were inconsistent, suggesting that they were previously misidentified. The tephra at 90–95 cm in core H211 was only tentatively identified¹¹ as Rotoma. A re-identification of this ash as Waiohau (also from the Okataina volcanic centre and of similar mineralogy to the Rotoma⁹) is based on a re-examination of the ferromagnesian assemblage and is in agreement with its stratigraphic position in the core. Similarly, the ash at 77–78 cm in H214 is re-identified as the Mamaku. It was originally identified as the Whakatane¹¹. A re-examination of the ferromagnesian assemblage indicates it is the Mamaku (of similar mineralogy to the Whakatane, and from the Okataina centre), which is consistent with its stratigraphic position in the core. The Rotoma is identified as the tephra lying below this ash in this core⁹ and is known to be present in this locality²⁰. If these two ashes are excluded from our discussion, it does not affect the reservoir calculations for any time slice.

* Dates on this tephra are out of sequence indicating disturbance in this core. These ages are not included in averages for this study.

† This data is younger than the terrestrial ¹⁴C age for the tephra indicating disturbance above the tephra. This data is not included in the average for this tephra.



ages (Table 3), indicating Holocene ocean circulation conditions similar to today. During the early deglaciation, at 14,700 $^{14}\text{Cyr BP}$ (Rerewhakaaitu tephra) the surface-reservoir age was 310 years, similar to modern values, within errors. Lower atmospheric CO_2 in the glaciation and early deglaciation would have decreased the ^{14}C of the ocean at this time⁶; thus the low reservoir age at 14,700 $^{14}\text{Cyr BP}$ indicates that the return to modern circulation by this time was strong enough to more than counter this effect.

At about 11,850 $^{14}\text{Cyr BP}$ (Waiohau tephra) we estimate a surface-reservoir age of 800 years (twice the modern value), a ventilation age at 1,675 m of 1,650 years (similar to the modern value), and an equivalent surface-to-deep-water age of 850 years (about 500 years younger than modern; Table 3, Fig. 3b). These results are consistent with previous benthic-planktonic foraminiferal ^{14}C differences indicating reduced benthic-planktonic ages relative to present at around 12,000 $^{14}\text{Cyr BP}$ (refs 5 and 13).

Our data from the Waiohau indicates that surface-reservoir ages in the subtropical Pacific were older at the time of the Antarctic Cold Reversal but that deep-reservoir ages were unchanged, indicating that shallow, not deep, circulation was affected. Ocean circulation changes associated with brief climate variations such as the Younger Dryas can alter atmospheric ^{14}C levels^{2–4,14}. At about 11,700 $^{14}\text{Cyr BP}$ a smaller, but similar change in atmospheric ^{14}C levels has been determined⁴. This is just after melt-water peak 1A and the re-initiation of NADW flux¹⁵, coincident with both the Antarctic Cold Reversal^{4,16,17} and the deposition of the Waiohau tephra. If the resulting increased ages of the subtropical surface water were not the result of atmospheric ^{14}C variations as a consequence of the deglaciation process⁴ they may have been associated with changes in subantarctic shallow-water mass¹⁸ production¹⁴. Older surface-reservoir ages could have been translated to South Pacific subtropical thermocline waters as a transient effect caused by increased leakage of subantarctic surface waters across the subtropical convergence.

During the last glaciation (Kawakawa tephra 22,590 $^{14}\text{Cyr BP}$)⁹ the

surface-reservoir age in the Bay of Plenty was 1,990 years. The Kawakawa tephra is also found south of the Chatham Rise¹⁰ (Fig. 1), allowing us to estimate the glacial reservoir ages of subpolar waters. Subpolar surface ages during the glaciation were 1,970 years, the same as regional subtropical waters. Modern subtropical and subpolar surface ages differ by 100–200 years (ref. 12). Thus, subtropical and subpolar surface ages were 4–5 times modern values (Table 1). Such large increases in subtropical surface-reservoir ages contradict box model calculations which indicate that subtropical surface-reservoir ages are insensitive to ocean circulation changes^{6,8}.

The Kawakawa tephra was associated with sufficient benthic foraminifera to obtain apparent ventilation ages from three cores at 1,300, 2,065 and 2,700 m water depth. Deep-reservoir ages increase with increasing depth yielding apparent ventilation ages of 3,000, 3,470 and 5,040 respectively, which are 1,500–3,000 years greater than today (Table 3; Fig. 3c). In contrast, owing to the old surface-reservoir ages, the benthic-planktonic (surface-to-deep-water) age of 1,030 in our shallowest core is similar to today (Table 3), and consistent with previous benthic-planktonic age studies, indicating that glacial radiocarbon age of Pacific deep waters were only 100–500 years greater than today^{5,8,13}. Significantly, this indicates that during the glaciation, although the surface-to-deep-water age contrast changed minimally, deep-water ages were 2–3 times larger than modern. Additionally, ventilation ages increase nearly twofold between 2,000 and 2,700 m (Fig. 3c).

The large ventilation ages and ^{14}C vertical structure during the glaciation bear directly on conflicting reconstructions of Southern Ocean deep-water palaeocirculation. Benthic foraminiferal $\delta^{13}\text{C}$ (refs 15,19,20) and carbonate dissolution²¹ indicate a gradient between 'older' poorly ventilated deep waters and better ventilated 'younger' intermediate waters. Authigenic uranium also suggests more poorly ventilated Southern Ocean deep waters²². In contrast, trace-metal indices in benthic foraminifera^{23,24} and U-series tracers²⁵ suggest similar rates of ventilation for the Holocene and glacial

Table 2 Modern samples

Location	Collection date	Sample	Lab Code	^{14}C age	Error
Waitangi Beach, Chatham Is.	1933	Gastropod 1*	CAMS 40758	690	40
Waitangi Beach, Chatham Is.	1933	Gastropod 2	CAMS 40759	540	40
Waitangi Beach, Chatham Is.	1933	Gastropod 3A†	CAMS 40760	550	40
Waitangi Beach, Chatham Is.	1933	Gastropod 3B†	CAMS 40761	560	50
Waitangi Beach, Chatham Is.	1933	Gastropod 4	CAMS 40852	590	40
Awanui Bay, East Cape, NZ	1924	Bivalve 1A‡	CAMS 40762	470	40
Awanui Bay, East Cape, NZ	1924	Bivalve 1B‡	CAMS 40763	460	40

* Gastropod 1 (CAMS 40758) was a fragment with bio-erosion, probably not alive at time of collection (not used in pooling for averages).

† 3A and 3B are sub-samples of the same gastropod.

‡ 1A and 1B are two values from one articulated bivalve.

Modern sample results presented in conventional ages. Reported ages have been age corrected for the date of collection, which for all samples is pre-bomb. Analyses were made at the Centre for Accelerator Mass Spectrometry (CAMS), Lawrence Livermore National Laboratory, California, USA.

Table 3 Tephra ages and ocean reservoir estimates

Location	Terrestrial ^{14}C age	Planktonic foraminiferal ^{14}C age	Surface reservoir age	Benthic foraminiferal ^{14}C age	Apparent ventilation age	Modern ventilation age	Past surface to deepwater age difference	Modern surface to deepwater age difference	Water depth (m)
Bay of Plenty									
Modern	0	470 ± 40	470 ± 40						
Whakatane	4,830 ± 20	5,190 ± 50	360 ± 50	6,350 ± 80	1,520 ± 80	~1,800	1,160 ± 90	~1,330	2,065 (H213)
Mamaku	7,250 ± 20	7,610 ± 60	360 ± 60						
Rotoma	8,530 ± 10	8,920 ± 40	390 ± 40						
Waiohau	11,850 ± 60	12,650 ± 90	800 ± 110	13,500 ± 60	1,650 ± 80	~1,500	850 ± 130	~1,030	1,675 (H209)
Rerewhakaaitu	14,700 ± 110	15,010 ± 110	310 ± 160						
Kawakawa	22,590 ± 230	24,580 ± 140	1,990 ± 270	26,060 ± 150	3,470 ± 270	~1,800	1,480 ± 210	~1,330	2,065 (H213)
Chatham Rise									
Modern	0	560 ± 40	560 ± 40	25,590 ± 380	3,000 ± 440	~1,300	1,030 ± 540	~740	1,300 (U939)
Kawakawa	22,590 ± 230	24,560 ± 320	1,970 ± 390	27,630 ± 210	5,040 ± 310	~1,400	3,070 ± 440	~840	2,700 (U938)

Summary of the terrestrial⁹ and marine ^{14}C ages of the tephra and ocean reservoir estimates in this study. The marine age for each tephra was obtained by: (1) forming weighted averages, using reported errors on AMS ^{14}C dates ($1/\sigma^2$) as weights, of the radiocarbon dates above and below the tephra respectively; (2) taking the mean of the weighted averages from above and below the tephra obtained in step (1), providing core-specific tephra ages; and (3) taking the mean of core-specific tephra ages for identical tephra. The error associated with the marine ^{14}C ages of tephra is a pooled estimate of the standard deviation using individual reported errors. Errors for reservoir ages, ventilation ages and surface-to-deep water ages, are calculated from the tephra uncertainties using gaussian error propagation. All error estimates in this table are rounded to the nearest 10-year interval.

Southern Ocean. Our results support longer residence times for Southern Ocean deep waters in the glaciation and are consistent with a strong gradient between 'older' poorly ventilated deep waters and better ventilated 'younger' intermediate waters.

The large reservoir ages we observe in the early glaciation are likely to represent short-lived transients, as CO_2 exchange dynamics^{6,14} tend to re-equilibrate atmospheric and oceanic ^{14}C reservoirs on short timescales⁶. An increase in ^{14}C production or

changes in atmosphere–ocean partitioning associated with large circulation changes could account for the enhanced gradient. However, cosmogenic production rates are an unlikely cause because estimates of ^{14}C production rates during the late glaciation are nearly constant^{26,27}. The existence of an atmospheric ^{14}C plateau near about 25,000 ^{14}C yr BP (ref. 28) and our vertical hydrographic ^{14}C profile (Fig. 3c) occurring shortly thereafter leads us to speculate that these may be connected by a significant change in ocean

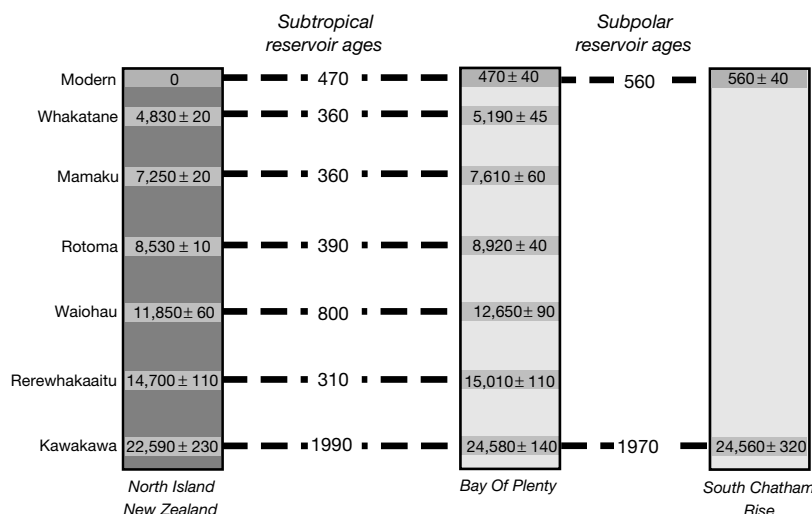


Figure 2 Surface-reservoir ages for subtropical and subpolar waters. During the Last Glacial Maximum surface-reservoir ages were ~2,000 years in both subtropical and subpolar surface waters. At 11,850 ^{14}C yr BP, surface-reservoir ages in subtropical waters were ~800 years. Comparison of marine and terrestrial ^{14}C ages for the Waiohau tephra in cores H209, H211 and H214 show significant variation between the three cores; the surface-reservoir age estimate at 11,850 ^{14}C yr BP is ~1,030 years in H214, ~840 years

in H209 and ~530 years in H211. The smaller surface-reservoir age estimate in H211 may be due to higher bioturbation intensity associated with lower sedimentation rates above the Waiohau tephra in that core (5.8 cm kyr⁻¹ above versus 25.3 cm kyr⁻¹ below). In cores H214 and H209 the ^{14}C ages above and below the tephra differ by no more than 100 years, whereas in core H211 the ages differ by ~550 years. We therefore consider the surface-reservoir age estimates from cores H209 and H214 to be more reliable.

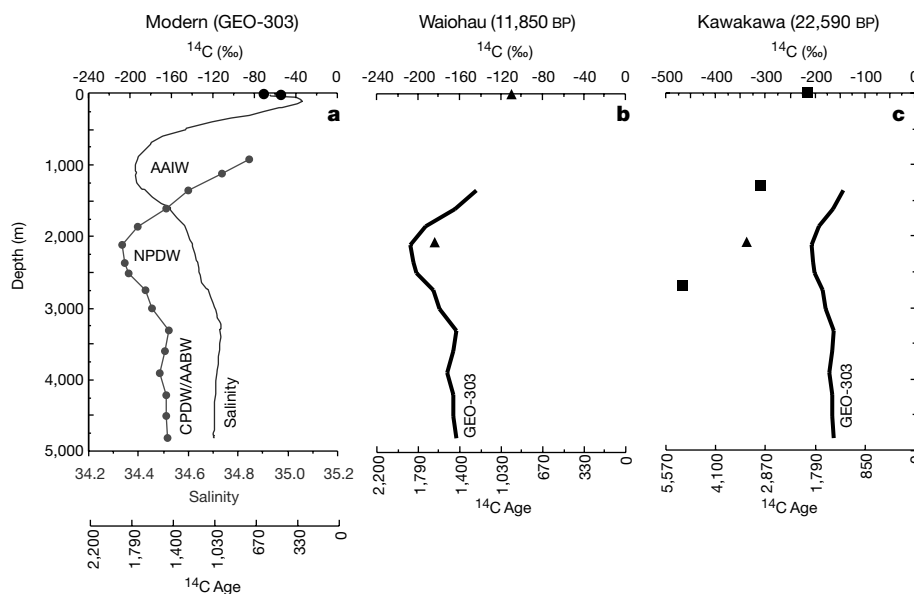


Figure 3 Surface and deep reservoir ages in the southwest Pacific for three time slices. **a**, Modern carbonate samples from the Bay of Plenty and Chatham Islands (solid circles). The modern profile is from GEOSECS station 303 (38° S, 170° W) and includes radiocarbon ($\Delta^{14}\text{C}$, and equivalent ^{14}C age; grey circles) and salinity (solid line). Major water masses are labelled: Antarctic Intermediate Water (AAIW), North Pacific Deep Water (NPDW), Antarctic Bottom Water (AABW) / Circumpolar Deep Water (CPDW). (We note that the horizontal scale changes for this modern profile, which is plotted with the palaeo- $\Delta^{14}\text{C}$ estimates for reference in **b** and **c**. **b**, Deglacial period, Waiohau (11,805 ^{14}C yr BP, solid

triangles). **c**, Last glaciation, Kawakawa (22,590 ^{14}C yr BP; Bay of Plenty core, solid triangles; Chatham Rise cores, solid squares). Palaeo- $\Delta^{14}\text{C}$ estimates were calculated using the respective foraminifera age relative to the atmosphere. During the Waiohau, $\Delta^{14}\text{C}$ at 1,675 m is similar to modern pre-bomb values, whereas the surface-reservoir age is older. During the Kawakawa, surface-reservoir ages and apparent ventilation ages are significantly older than today. The sharp increase in apparent ventilation ages with water depth supports interpretations of a strong gradient between 'older' poorly ventilated deep waters and 'younger' better ventilated intermediate waters.

circulation during the onset of full glaciation. A reduction in deep-ocean ventilation before about 25,000 ^{14}C yr BP as the climate system came into full glacial conditions (at or about the stage 3/2 boundary), followed by a return to more modestly reduced production later in the glacial period could explain the large apparent ventilation age, acknowledging that the exact timing of the ^{14}C plateau²⁸ is somewhat uncertain. Such a transient effect would have produced older deep water that continued to work its way through the deep ocean at about 22,590 ^{14}C yr BP.

Our results have important implications for marine ^{14}C chronologies which assume a constant reservoir correction. For example, greater-than-modern surface-reservoir ages in the Southern Hemisphere would cause synchronous events to appear artificially earlier in the south than elsewhere. Our data provide constraints for models of the radiocarbon cycle which seek to explain atmospheric ^{14}C changes in terms of ocean dynamics^{4,28}, provide diagnostics of glacial thermohaline circulation, and address mechanisms for reduced glacial atmospheric $p\text{CO}_2$. □

Received 9 August 1999; accepted 31 March 2000.

1. Stuiver, M. & Polach, H. A. Discussion and reporting of ^{14}C data. *Radiocarbon* **19**, 355–363 (1977).
2. Bard, E. *et al.* The North Atlantic atmosphere-sea surface ^{14}C gradient during the Younger Dryas climatic event. *Earth Planet. Sci. Lett.* **126**, 275–287 (1994).
3. Edwards, R. L. *et al.* A large drop in atmospheric $^{14}\text{C}/^{12}\text{C}$ and reduced melting in the Younger Dryas, documented with ^{230}Th ages of corals. *Science* **260**, 962–968 (1993).
4. Hughen, K. A., Overpeck, J. T., Peterson, L. C. & Trumbore, S. Rapid climate changes in the tropical Atlantic during the last deglaciation. *Nature* **380**, 51–54 (1996).
5. Adkins, J. F. & Boyle, E. A. Changing atmospheric $\Delta^{14}\text{C}$ and the record of deep water paleoventilation ages. *Paleoceanography* **12**, 337–344 (1997).
6. Bard, E. Correction of accelerator mass spectrometry ^{14}C ages measured in planktonic foraminifera: Paleoceanographic implications. *Paleoceanography* **3**, 635–646 (1988).
7. Östlund, G. & Stuiver, M. GEOSECS Pacific radiocarbon. *Radiocarbon* **22**, 25–53 (1980).
8. Broecker, W. S., Peng, T.-H., Trumbore, S., Bonani, G. & Wolfli, W. The distribution of radiocarbon in the glacial ocean. *Glob. Biogeochem. Cycles* **4**, 103–117 (1990).
9. Froggatt, P. C. & Lowe, D. J. A review of late Quaternary silicic and some other tephra formations from New Zealand: their stratigraphy, nomenclature, distribution, volume, and age. *NZ J. Geol. Geophys.* **33**, 89–109 (1990).
10. Carter, L., Nelson, C. S., Neil, H. L. & Froggatt, P. C. Correlation, dispersal, and preservation of the Kawakawa Tephra and other late Quaternary tephra layers in the Southwest Pacific Ocean. *NZ J. Geol. Geophys.* **38**, 29–46 (1995).
11. Kohn, B. P. & Glasby, G. P. Tephra distribution and sedimentation rates in the Bay of Plenty, New Zealand. *NZ J. Geol. Geophys.* **21**, 49–79 (1978).
12. Higham, T. F. G. & Hogg, A. G. Radiocarbon dating of prehistoric shell from New Zealand and calculation of the ΔR value using fish otoliths. *Radiocarbon* **37**, 409–416 (1996).
13. Shackleton, N. J. *et al.* Radiocarbon age of last glacial Pacific deep water. *Nature* **335**, 708–711 (1988).
14. Stocker, T. F. & Wright, D. G. Rapid changes in ocean circulation and atmospheric radiocarbon. *Paleoceanography* **11**, 773–795 (1996).
15. Charles, C. D. & Fairbanks, R. G. Evidence from Southern Ocean sediments for the effect of North Atlantic deep-water flux on climate. *Nature* **355**, 416–419 (1992).
16. Blunier, T. *et al.* Timing of the Antarctic cold reversal with respect to the Younger Dryas event. *Geophys. Res. Lett.* **24**, 2683–2686 (1997).
17. Sowers, T. & Bender, M. Climate records covering the last deglaciation. *Science* **269**, 210–213 (1995).
18. Toggweiler, J. R., Dixon, K. & Broecker, W. S. The Peru upwelling and the ventilation of the South Pacific thermocline. *J. Geophys. Res.* **C 11**, 20467–20497 (1991).
19. Duplessy, J.-C. *et al.* Deepwater source variations during the last climatic cycle and their impact on the global deepwater circulation. *Paleoceanography* **3**, 343–360 (1988).
20. McCorkle, D. C., Hogg, D. T. & Veeh, H. H. Glacial and Holocene stable isotopic distributions in the southeastern Indian Ocean. *Paleoceanography* **13**, 20–34 (1998).
21. Howard, W. R. & Prell, W. L. Late Quaternary carbonate production and preservation in the Southern Ocean: Implications for oceanic and atmospheric carbon cycling. *Paleoceanography* **9**, 453–482 (1994).
22. Francois, R. *et al.* Contribution of Southern Ocean surface-water stratification to low atmospheric CO_2 concentrations during the last glacial period. *Nature* **389**, 929–935 (1997).
23. Rosenthal, Y., Boyle, E. A. & Labeyrie, L. D. Last glacial maximum paleochemistry and deepwater circulation in the Southern Ocean: Evidence from foraminiferal cadmium. *Paleoceanography* **12**, 787–796 (1997).
24. Lea, D. W. A trace metal perspective on the evolution of Antarctic Circumpolar Deep Water chemistry. *Paleoceanography* **10**, 733–748 (1995).
25. Yu, E.-F., Francois, R. & Bacon, M. P. Similar rates of modern and last-glacial thermohaline circulation inferred from radiocarbon data. *Nature* **379**, 689–694 (1996).
26. Finkel, R. C. & Nishizumi, K. Beryllium 10 concentrations in the Greenland Ice Sheet Project 2 ice core from 3–40 ka. *J. Geophys. Res.* **102**, 26699–26706 (1997).
27. Guyodo, Y. & Valet, J.-P. Relative variations in geomagnetic intensity from sedimentary records: the past 200,000 years. *Earth Planet. Sci. Lett.* **143**, 23 (1996).
28. Kitagawa, H. & van der Plicht, J. A 40,000-year varve chronology from Lake Suigetsu, Japan: Extension of the ^{14}C calibration curve. *Radiocarbon* **40**, 505–515 (1998).
29. Fairbanks, R. G., Wiebe, P. H. & Bé, A. W. H. Vertical distribution and isotopic composition of living planktonic foraminifera in the western North Atlantic. *Science* **207**, 61–63 (1979).
30. Wright, L. C., McGlone, M. S., Nelson, C. S. & Pillans, B. J. An integrated latest Quaternary (Stage 3 to present) paleoclimatic and paleoceanographic record from offshore northern New Zealand. *Quat. Res.* **44**, 283–293 (1995).

31. Weaver, P. P. E., Neil, H. L. & Carter, L. Sea surface temperature estimates from the Southwest Pacific based on planktonic foraminifera and oxygen isotopes. *Paleogeogr. Paleoclimatol. Palaeoecol.* **131**, 241–256 (1997).

Acknowledgements

We thank E. Bard and J. Southon for discussions and comments. L. Carter kindly provided material from the core collection and modern shells were provided by W. Blom. L. Robertson helped produce the figures and P. Shane helped with tephra identifications. Radiocarbon analyses at CAMS were performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory.

Correspondence and requests for materials should be addressed to E. L. S. (e-mail: e.sikes@auckland.ac.nz).

Multiple seismic discontinuities near the base of the transition zone in the Earth's mantle

N. A. Simmons & H. Gurrila

Department of Geosciences, Texas Tech University, Lubbock, Texas 79409, USA

The seismologically defined boundary between the transition zone in the Earth's mantle (410–660 km depth) and the underlying lower mantle is generally interpreted to result from the breakdown of the γ -spinel phase of olivine¹ to magnesium-perovskite and magnesiowüstite². Laboratory measurements of these transformations of olivine have determined that the phase boundary has a negative Clapeyron slope and does indeed occur near pressures corresponding to the base of the transition zone^{2,3}. But a computational study has indicated that, because of the presence of garnet minerals, multiple seismic discontinuities might exist near a depth of 660 km (ref. 4), which would alter the simple negative correlation of changes in temperature with changes in the depth of the phase boundary. In particular, garnet minerals undergo exothermic transformations near this depth, acting to complicate the phase relations^{5–9} and possibly effecting mantle convection processes in some regions⁹. Here we present seismic evidence that supports the existence of such multiple transitions near a depth of 660 km beneath southern California. The observations are consistent with having been generated by garnet transformations coupling with the dissociation of the γ -spinel phase of olivine. Temperature anomalies calculated from the imaged discontinuity depths—using Clapeyron slopes determined for the various transformations⁴—generally match those predicted from an independent P-wave velocity model of the region.

Non-olivine components in upper mantle chemistry may be significant in the nature of the transition zone and mantle convection^{4,10}. Ringwood¹⁰ suggests that subduction materials are likely to accumulate 660 km below delamination and transformation of former oceanic crust. This process may generate a 'garnetite' layer near a depth of 660 km, which possibly impedes penetration of subducted oceanic lithosphere. This layering may effectively hinder whole-mantle convection at some locations. In Ringwood's model, relatively young and thin oceanic slabs may be too buoyant to penetrate the pre-existing garnetite layer and may therefore be trapped above the 660-km discontinuity, becoming part of the transition zone. This implies that lateral heterogeneity in transition-zone chemistry might occur. These variations in mantle chemistry would most probably cause lateral variations in velocity contrasts across the base of the transition zone.

The negative Clapeyron slope of the olivine system would have the effect of deepening the transition boundary in cold regions (such as subduction zones) and shallowing the transition boundary in hot regions. The positive Clapeyron slope of the proposed garnet system would, however, have the opposite effect. Combined, these two transformations could create a situation in which the perovskite-forming boundary could deepen in hot as well as cold regions, providing no absolute discrepancy between temperature regimes⁹. In colder regions, however, it may be possible to detect a distinct precursory garnet-to-ilmenite transformation, which could provide some thermal discrepancy⁴. In either case, the relative dominance of the garnet and olivine systems is a principal variable in determining the true depth of the base of the transition zone.

Southern California offers an excellent laboratory for seismic investigations of the upper mantle transition zone. This region has the densest collection of broadband seismic stations available through the Incorporated Research Institutions for Seismology (IRIS) Data Management Center, and southern California is ideally located to record a large number of teleseismic events from the seismically active Pacific Rim. The azimuthal coverage of seismic events recorded in California is therefore sufficient to generate a three-dimensional (3D) image of the transition zone. The distribution of S-wave piercing points at 660 km is illustrated in Fig. 1.

We applied a form of common conversion point stacking^{11–13} of Pds (P-phase converted to S at a depth *d*) and PPds (similar to Pds but for an incoming PP phase) receiver functions^{14,15} and converted to the depth domain using the IASPI91 one-dimensional (1D) global reference velocity model¹⁶ with local 3D velocity corrections using the NA95 North American velocity model¹⁷. This process uses variable bin sizes to enhance small-scale features in dense data areas while expanding stacking radii in areas with lesser data; therefore, adjustment of stacking criteria within this algorithm allows us to enhance fine discontinuity structure, where possible, while retain-

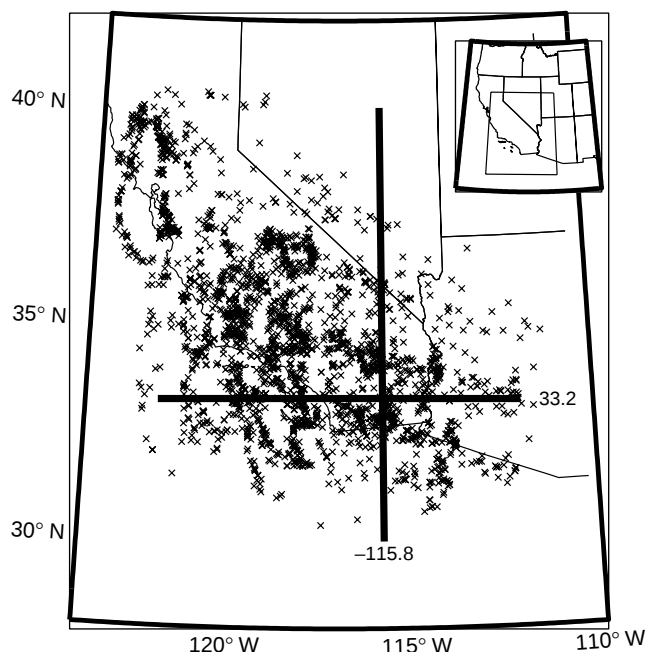


Figure 1 P660s and PP660s conversion points for 2,609 receiver functions used in this study. This map includes only those conversion points from events that produced visibly stable receiver functions using a 0.3-Hz gaussian filter. Conversion points were determined by back ray tracing through the NA95 3D velocity model¹⁷. Locations of selected cross-sections transecting the Salton Trough region are also plotted and labelled according to their respective line of latitude or longitude. See Fig. 3 for these cross-sections.

ing lateral continuity in areas with smaller amounts of data. This method has proved to be extremely effective, producing a high-resolution image that we propose will confirm the existence of multiple discontinuities at the base of the transition zone beneath southern California.

Figure 2 shows the seismic signature of stacked receiver functions with a range of different deconvolution frequencies. The relative signal strengths due to garnet transformations, determined by theoretical velocity contrasts, are small compared with the velocity contrast across the entire zone⁴, making them undetectable in areas with poor data coverage or by methods using low frequency data (0.1 Hz, Fig. 2a). However, the garnet-to-perovskite transformation is sufficiently abrupt to generate an observable seismic signal¹⁸, which can be seen in certain localities in our high-frequency image (0.3 Hz, Fig. 2c). At this frequency, the base of the transition zone appears to be affected by the olivine and the garnet systems. This particular location may be representative of a low-temperature profile in which the precursory garnet-to-ilmenite transformation becomes separable from the olivine dissociation pulse that is only visible at higher frequencies. Synthetic analysis shown in Fig. 2d demonstrates the scale of seismic signal expected, with conservative estimates of the probable velocity and density contrasts for the

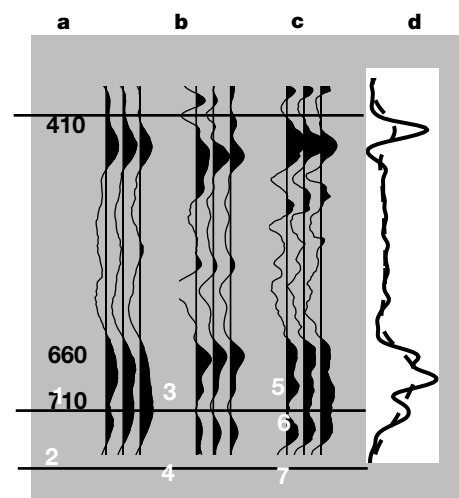


Figure 2 Stacked receiver functions centred at 34.4, 34.5 and 34.6°N computed for three different frequencies. These stacks were produced using a constant search radius of 0.4°. The average number of receiver functions in these stacks is 70, but all of these stacks include at least 25 events. The number of events varies amongst different stacks owing to variations in the stability of the deconvolution of receiver functions at the different frequencies (0.1, 0.2 and 0.3 Hz). This location is beneath the western Transverse Ranges, which demonstrates coherence of three discontinuities with increasing frequency. **a**, At low frequency (0.1 Hz), the multiple peaks appear to smear into a single peak. Location 1 appears to be combined effects of garnet-to-ilmenite and olivine dissociation, whereas location 2 may be the result of the garnet-to-perovskite transformation. **b**, These phase transformations appear as distinct peaks at 0.2 Hz, but locations 3 and 4 are similar to 1 and 2. **c**, At higher frequency (0.3 Hz), three distinct transformations are visible (locations 5–7). Location 5 is a shallow and separate discontinuity, possibly a result of a garnet-to-ilmenite transition, whereas peaks 6 and 7 are olivine dissociation and garnet-to-perovskite transformations, respectively. **d**, A synthetic example of the expected seismic signature for these three discontinuities. The synthetic data are stacked receiver functions computed for 0.1 Hz (dashed) and 0.3 Hz (solid) with a similar ray path distribution as the real receiver functions used in this figure. The velocity contrasts are conservative values based on published results⁴. Notice the similarity between the synthetics and the real stacks in **a** and **c**. **d** is intended to illustrate the expected amplitudes of Pds phases associated with the proposed phase changes. Because there are significant lateral variations in the depth to these discontinuities, we did not attempt to line up the depths in the synthetics with **a–c**.

proposed garnet and olivine transformations⁴. Observable seismic signal is expected from the olivine and garnet systems at the base of the transition zone when 0.3-Hz receiver functions are stacked (compare the 0.1-Hz stack, Fig. 2d, dashed line). With the number of events included in these stacks, the main source of noise is signal-generated owing to reverberations or scattering. Scattered energy is by definition randomized, but to incorporate possible errors caused by reverberations, the stacked receiver functions shown in Fig. 2d were computed from synthetics using the same distribution of ray parameters as those events contributing to the stacked receiver functions in Fig. 2a–c.

Several conversions near a depth of 660 km are also apparent in real stacks beneath the Salton Trough region (Fig. 3). The distance between the proposed precursory garnet transformation and the

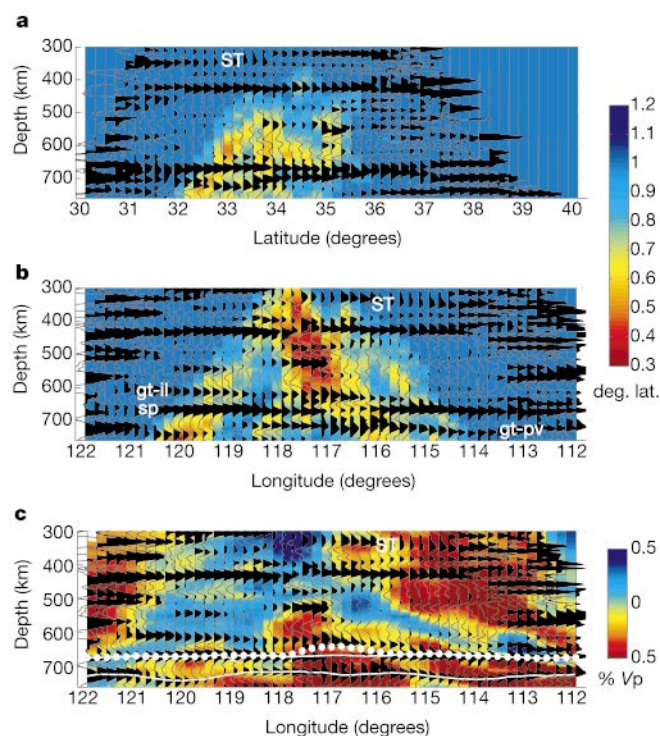


Figure 3 Representative portions of the 3D image beneath southern California showing multiple discontinuities near 660-km depth. These are 0.3-Hz receiver functions stacked with the requirement that 10 stations contribute to each value while not allowing the search radii to exceed 1° latitude. **a**, A north–south profile at 115.8°W longitude that cuts through the Salton Trough region (ST). The background of this section is the matrix of search radii (blue indicates a 1° search radius and a failure to include data from 10 different stations in a single stacking bin). **b**, An east–west profile through the Salton Trough region with the matrix of search radii in the background. Note that three distinct discontinuities appear near a 660-km depth. **c**, The same section as **b** with velocity perturbation (%) in the background and interpreted discontinuity horizons mapped. The suspected precursory garnet-to-ilmenite (gt-il) transformation near 610 km intersects the spinel transformation near 650 km beneath the Salton Trough region, suggestive of a high-temperature trend in a mixed garnet–olivine system. The section exhibits a deepening of garnet phases above and below the spinel system transformation (sp) in three localities. The garnet-to-perovskite transformation (gt-pv) appears near 720 km in this entire section. Note the anticorrelated behaviour of the sp and gt-pv discontinuities and their increased separation in the low-velocity regions which increase the BTT (see text). Solid lines are picked horizons assuming maximum lateral continuity near the centre of the section. Analysis of neighbouring profiles indicates that the weak phases on in this interpretation are smearing due to bin size. Stacks using a fixed, smaller bin size indicate that the shallower, stronger phases are more likely to be the 660-km discontinuity. Circles represent an alternative, preferred interpretation of the 660-km horizon, which has a significant structure near the centre low-velocity zone (near –117°). V_p is P-wave velocity.

660-km discontinuity diminishes from west to east (Fig. 3b, c, near longitude 115°W). At this location the suspected garnet-to-perovskite transition (~720 km) deepens from ~720 km to ~750 km, whereas the olivine dissociation shallows from ~680 km to ~660 km. This is consistent with what is expected given the anticorrelated Clapeyron slopes for these proposed transformations and a high-temperature anomaly at the base of the transition zone. An independent tomography model based on P-wave delay time inversion (K. Dueker, personal communication) supports the interpretation that a high-temperature anomaly is associated with the Salton Trough region near 33°N (Fig. 3c), as suggested by a 1% decrease in P-wave velocity imaged at the base of the transition zone. A north–south cross-section through the 3D image is shown in Fig. 3a; the two deepest phases appear anticorrelated but there is no strong evidence of a precursory ilmenite phase, which suggests that the signal strength of this phase is more variable than that of the deeper 720-km discontinuity.

To test the likelihood that the observed structures are thermally controlled, estimated thermal anomalies were calculated on the basis of velocity model perturbation and compared with those calculated from conservative estimates of the opposing Clapeyron slopes^{19,25}. For this comparison, we focused on the 660-km discontinuity and the garnet-to-perovskite transition near 720 km depth. We refer to the distance between these two phase changes as the basal transition thickness (BTT). Estimates of BTT are more accurate than those of absolute depth because variations in depth arising from differences in velocity models will affect the depths imaged for the 660-km and 720-km discontinuities to a similar extent. As a result, any errors due to lateral velocity variations in the velocity model (NA95) are minimized when differential structure on each discontinuity is used rather than absolute depths to the discontinuities²⁰. The results of this comparison are shown in Fig. 4.

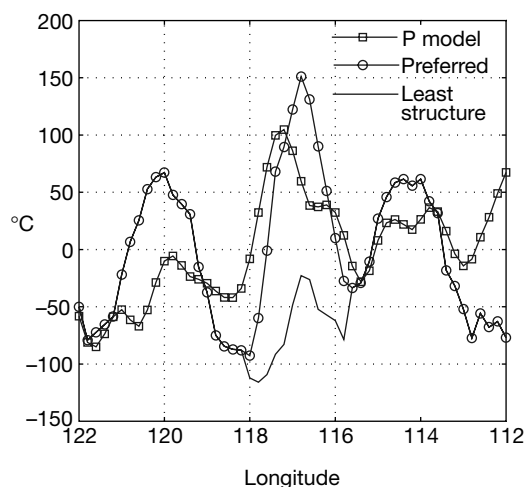


Figure 4 Calculated temperature anomalies for the cross-section at 33.2°N latitude (Fig. 3b, c). Squares represent temperature anomalies calculated from the average velocity perturbation between the two discontinuities. This is a simple conversion based on the equation $\partial V_p = (4.1 \pm 1.1 \times 10^{-4} \text{ km s}^{-1} \text{ °C}^{-1}) \partial T$ derived from relationships given by Anderson²⁴. The line without symbols represents temperature anomalies based on the distance between the BTT from the more conservative interpretation (white lines in Fig. 3c). Using Clapeyron slopes of –2.0 and 3.0 Mpa K^{–1} for the spinel dissociation and garnet-to-perovskite transition, respectively, yields a temperature-equation anomaly of $\partial T = \text{BTT}/0.15 \text{ °C}$. These two curves are similar in both magnitude and the location of most thermal anomalies, but fail to match the large thermal anomaly in the centre of the cross-section of the velocity model. Circles represent the temperature anomalies based on the preferred interpretation in Fig. 3c. The relative temperature range and trends calculated from this BTT are a better match to those derived from the velocity model, which supports our preferred interpretation.

The 250 °C range in temperature inferred from measurements of BTT is similar in magnitude to the range calculated from variations in the independent P-wave velocity model. The general trend in the locations of the peak thermal anomalies derived from both methods correlate well.

Computation of seismic velocity profiles from mineral physics data suggests that multiple transitions are present, and are most distinct at low temperatures such as those associated with subduction zones⁴. The additional boundaries are due to the garnet transformations to ilmenite (shallower than 660 km) and perovskite (deeper than 660 km). At subduction temperatures, mineral physics calculations suggest that distinct transitions occur at 608–664, 690–693 and 709–731 km. The existence of these three discontinuities has also been suggested by a seismic investigation of the subduction zone beneath northeast China where three distinct discontinuities between 660 and 780 km are found at the tip of the subducting slab²¹. Our interpreted depth variation from 730 to 750 km of the proposed garnet phase change requires almost one-third of the temperature anomaly suggested for the slab beneath China. Other seismic investigations found seismic signatures near the 660-km depth that were inconsistent with a simple transformation in a simple olivine system^{13,22,23}, which suggests that the base of the transition zone is not simply a dissociation of γ -spinel.

Southern California has been exposed to a spectrum of tectonic events which has created a thermally complex region to study (characterized by the recent subduction of a relatively cold oceanic slab followed by the subduction of a warm spreading ridge with possible continued down welling of cold upper mantle material along the transverse ranges). Observed variations in the BTT between the 660 and 720 km discontinuities (associated with the dissociation of olivine and garnet phases, respectively) beneath this thermally complex region supports the hypothesis that significant garnet transformations exist near the base of the upper mantle transition zone. The anticorrelated depth variations of these discontinuities are in agreement with the garnet–olivine systems proposed near a depth of 660 km. Calculations of the thermal anomalies based on the observed BTT are similar to those derived from an independent tomography model. These observations suggest that garnet phases must be considered when studying the base of the transition zone in thermally complex regions such as southern California and in regions with subducted slabs²¹. To determine the relative significance of garnet phase transformations coupled with olivine transformations, as global processes, will require additional work in less complex regions. □

Received 12 November 1999; accepted 10 April 2000.

1. Ringwood, A. E. in *Advances in Earth Sciences* (ed. Hurley, P. M.) 287–356 (MIT Press, Cambridge, Massachusetts, 1966).
2. Ito, E. & Takahashi, E. Postspinel transformations in the system Mg_2SiO_4 – Fe_2SiO_4 and some geophysical implications. *J. Geophys. Res.* **94**, 10637–10646 (1989).
3. Ito, E., Akaogi, M., Topor, L. & Navrotsky, A. Negative pressure-temperature slopes for reactions forming $MgSiO_3$ perovskite from calorimetry. *Science* **249**, 1275–1278 (1990).
4. Vacher, P., Mocquet, A. & Sotin, C. Computation of seismic profiles from mineral physics: the importance of the non-olivine components for explaining the 660 km depth discontinuity. *Phys. Earth Planet. Inter.* **106**, 275–298 (1998).
5. Herzberg, C. & Gasparik, T. Garnet and pyroxenes in the mantle: a test of the majorite fractionation hypothesis. *J. Geophys. Res.* **96**, 16263–16274 (1991).
6. Gasparik, T. Enstatite-jadeite join and its role in the Earth's mantle. *Contrib. Mineral. Petrol.* **111**, 283–298 (1992).
7. Gasparik, T. Melting experiments on the enstatite-diopside join at 70–224 kbar, including the melting of diopside. *Contrib. Mineral. Petrol.* **124**, 139–153 (1996).
8. Gasparik, T. A model for the layered upper mantle. *Phys. Earth Planet. Inter.* **100**, 197–212 (1997).
9. Weidner, D. J. & Wang, Y. Chemical- and Clapeyron-induced buoyancy at the 660 km discontinuity. *J. Geophys. Res.* **103**, 7431–7441 (1998).
10. Ringwood, A. E. Role of the transition zone and 660 km discontinuity in mantle dynamics. *Phys. Earth Planet. Inter.* **86**, 5–24 (1994).
11. Dueker, K. G. & Sheehan, A. F. Mantle discontinuity structure beneath the Colorado Rocky Mountains and High Plains. *J. Geophys. Res.* **103**, 7153–7169 (1998).
12. Li, A., Fisher, K. M., Wyssession, M. E. & Clarke, T. J. Mantle discontinuity and temperature under the North American continental keel. *Nature* **395**, 160–163 (1998).
13. Owens, T. J., Nyblade, A. A., Gurrrola, H. & Langston, C. A. 410 and 660 km discontinuity structure beneath Tanzania, East Africa. *Geophys. Res. Lett.* **27**, 827–830 (2000).

14. Owens, T. J., Zandt, G. & Taylor, S. R. Seismic evidence for an ancient rift beneath the Cumberland Plateau, Tennessee: a detailed analysis of broadband teleseismic P-waveforms. *J. Geophys. Res.* **89**, 7783–7795 (1984).
15. Ammon, C. J. The isolation of receiver effects from teleseismic P waveforms. *Bull. Seismol. Soc. Am.* **81**, 2504–2510 (1991).
16. Kennett, B. L. N. & Engdahl, E. R. Traveltimes for global earthquake location and phase identification. *Geophys. J. Int.* **105**, 429–465 (1991).
17. Van der Lee, S. & Nolet, G. Upper mantle S-velocity structure of North America. *J. Geophys. Res.* **102**, 22815–22838 (1997).
18. Stixrude, L. Structure and sharpness of phase transitions and mantle discontinuities. *J. Geophys. Res.* **102**, 14835–14852 (1997).
19. Weidner, D. J. & Wang, Y. Chemical- and Clapeyron-induced buoyancy at the 660 km discontinuity. *J. Geophys. Res.* **103**, 7431–7441 (1998).
20. Gurrrola, H. & Minster, J. B. Thickness estimates of the upper-mantle transition zone from bootstrapped velocity spectrum stacks of receiver functions. *Geophys. J. Int.* **133**, 31–43 (1998).
21. Niu, F. & Kawakatsu, H. Complex structure of the mantle discontinuities at the tip of the subducting slab beneath the Northeast China: a preliminary investigation of broadband receiver functions. *J. Phys. Earth* **44**, 701–711 (1996).
22. Revenaugh, J. & Jordan, T. H. Mantle layering from ScS reverberations: 2. The transition zone. *J. Geophys. Res.* **96**, 19763–19780 (1991).
23. Shearer, P. M. & Flanagan, M. P. Seismic velocity and density jumps across the 410- and 660-kilometer discontinuities. *Science* **285**, 1545–1548 (1999).
24. Anderson, D. L. *Theory of the Earth* (Blackwell Science, Oxford, 1989).
25. Bina, C. R. & Helffrich, G. Phase transition Clapeyron slopes and transition zone seismic discontinuities. *J. Geophys. Res.* **99**, 15853–15860 (1994).

Acknowledgements

We thank the IRIS Data Management Center for data availability and S. van der Lee for expedient delivery and usage of the NA95 velocity model, and K. Dueker for allowing us to use his western USA tomography model before publication. This work was partially supported by the NSF.

Correspondence and requests for materials should be addressed to H.G.

Are lemmings prey or predators?

P. Turchin*, L. Oksanen†, P. Ekerholm‡, T. Oksanen‡ & H. Henttonen§

* Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269, USA

† Department of Ecological Botany, Umeå University, S-901 87 Umeå, Sweden

‡ Department of Animal Ecology, Umeå University, S-901 87 Umeå, Sweden

§ Finnish Forest Research Institute, Vantaa Research Center, PO Box 18, FIN-01301 Vantaa, Finland

Large oscillations in the populations of Norwegian lemmings have mystified both professional ecologists and lay public^{1–3}. Ecologists suspect that these oscillations are driven by a trophic mechanism^{4,5}: either an interaction between lemmings and their food supply, or an interaction between lemmings and their predators. If lemming cycles are indeed driven by a trophic interaction, can we tell whether lemmings act as the resource ('prey') or the consumer ('predator')? In trophic interaction models, peaks of resource density generally have a blunt, rounded shape, whereas peaks of consumer density are sharp and angular. Here we have applied several statistical tests to three lemming datasets and contrasted them with comparable data for cyclic voles. We find that vole peaks are blunt, consistent with their cycles being driven by the interaction with predators. In contrast, the shape of lemming peaks is consistent with the hypothesis that lemmings are functional predators, that is, their cycles are driven by their interaction with food plants. Our findings suggest that a single mechanism, such as interaction between rodents and predators, is unlikely to provide the 'universal' explanation of all cyclic rodent dynamics.

Many models of resource–consumer interactions predict that population cycles of the two species will be characterized by distinct shapes: blunt rounded peaks for resources, and sharp angular peaks in the consumer density (Fig. 1). This general pattern arises in

predator–prey^{6–8} and host–parasitoid⁹ models, but is particularly prominent in models developed specifically for small rodents^{10–13} for the following reason. Models of interaction between rodents and their specialized predators, weasels^{10–12}, must include a self-limitation term in the prey equation (without rodent self-limitation no cycle can occur because the rodent population will escape predator control by virtue of their much higher rate of population growth). A typical rodent–predator cycle occurs as follows. Starting from low prey and predator numbers, prey density rapidly builds up to the point where its further growth is prevented by self-limitation mechanisms, such as social interactions. Predators, on the other hand, can start increasing only after prey density grows beyond the threshold where predators can maintain positive energy balance (this threshold is related to the half-saturation constant in the predator functional response). By the time predators have increased to the point where they cause prey to decline, prey population has spent a prolonged period of time at peak densities. As a result, the shape of prey peaks will be blunt. By contrast, predator population grows exponentially while prey density is above the threshold, followed by rapid declines caused by starvation (or emigration). As a result, predator trajectories are characterized by ‘saw-shaped’ dynamics with sharp peaks. Models in which rodents are consumers^{10,13} behave in essentially the same manner, except that it is now rodent trajectories that are characterized by sharp peaks.

We emphasize that distinctive topological features of prey versus predator cycles are a generic feature of two-species trophic models (and are robust with respect to adding stochasticity). Predator peaks must be sharp: we cannot ‘flatten’ them by adding predator self-limitation, because this leads to a loss of cycles (stabilization of the

system). The bluntness of prey peaks is more variable, as the length that prey spent at the upper density threshold will depend on the relative growth rates of prey and predators. If prey growth rate is similar to (or slower than) that of predators, it may be difficult in practice to detect the plateau phase in prey dynamics. Because rodents are characterized by much faster reproductive rates than their predators, however, peak topology should be a particularly useful diagnostic for their dynamics.

Rodent–vegetation models have been applied to lemmings¹³, but unlike the vole case, we still lack good parameter estimates for these models, or manipulative experiments to test rival hypotheses. The theoretical observations discussed above, however, allow us to design an empirical test to distinguish between the two rival hypotheses for lemming cycles: one invoking the interaction with the food supply, and the other with predators. To pursue this idea, we located all time series data on the population dynamics of Norwegian lemmings (*Lemmus lemmus*) that were at least 20 years in length and had at least 2 observations per year (see Methods). We also analysed a comparable set of three vole series from the northern-most part of Fennoscandia (latitude $\geq 68^\circ\text{N}$), where vole populations exhibit high-amplitude oscillations.

Visual examination of the data plotted on logarithmic scale suggests a striking difference between the lemming and vole time series (Fig. 2). Lemmings have very sharp peaks, with rarely more than one observation period at the peak, whereas vole populations spend ~ 2 years in the vicinity of maximum densities before their populations collapse. One way to test statistically whether this difference is real is to examine the frequency distribution of log-transformed densities. The distribution of prey densities should be characterized by a negative skewness (a few observations at very low density, but most near the maximum density). Predator distribution, on the other hand, should be symmetric (no skewness). Indeed, there is a statistically significant difference between skewness of lemming and vole data (Table 1). Another way to look at the same issue is to compute the annual rate of population increase preceding peak density, $r_{\text{pre-peak}}$, which should be near zero for the resource, but much greater than zero for the consumer population. This is precisely the pattern that we see (Table 1). These results are therefore consistent with the hypothesis that lemmings are functional predators, whereas voles are functional prey.

An additional feature of numerical dynamics, the variability of peak densities, provides a further clue about dynamical mechanisms that may be responsible for lemming and vole cycles. Prey should have rather stereotypical dynamics at the population peak because they hit the population ceiling imposed by density-dependent regulation. In contrast, predator dynamics do not have a comparable ‘hard ceiling’ because predator density will start to decline when predators run out of food, and when this occurs depends very much on the timing of predator increase with respect to seasonality. Consider a threshold predator density, $N_{\text{threshold}}$, above which predators are expected to run out of winter food supply and

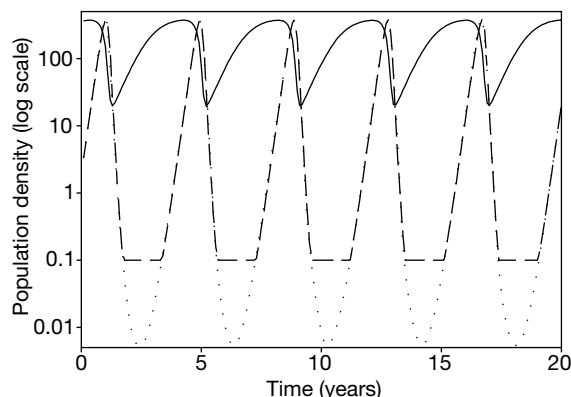


Figure 1 Shape of predator and prey peaks in a theoretical model⁶. Solid curve shows prey density. Dotted curve shows predator density. Broken line is the predator trajectory that would be observed when there is a threshold below which predator density is undetectable. Note the log scale of the y axis (thus, exponential growth/decline periods are represented by linear segments in the curves).

Table 1 Results of statistical analyses

	Skewness			Peak characteristics		
	Log-transformed	Log-transformed, zeroes omitted	Not log-transformed	Mean $r_{\text{pre-peak}}$	n	CV of peak density
Lemmings						
Finse	0.33	0.40	3.80	3.420	6	1.360
Kilpisjärvi	1.44	-0.15	4.15	2.966	4	0.805
Finnmark	0.76	0.96	4.32	3.030	2	0.845
Voies						
Kilpisjärvi	-0.80	-0.50	1.21	0.593	9	0.311
Pallasjärvi	-1.18	-0.56	1.56	0.361	4	0.430
Finnmark	-0.57	-0.57	1.38	0.692	4	0.339
t -test statistic	3.412	2.948	14.701	15.016		3.529
P	0.027	0.042	0.0001	0.0001		0.024

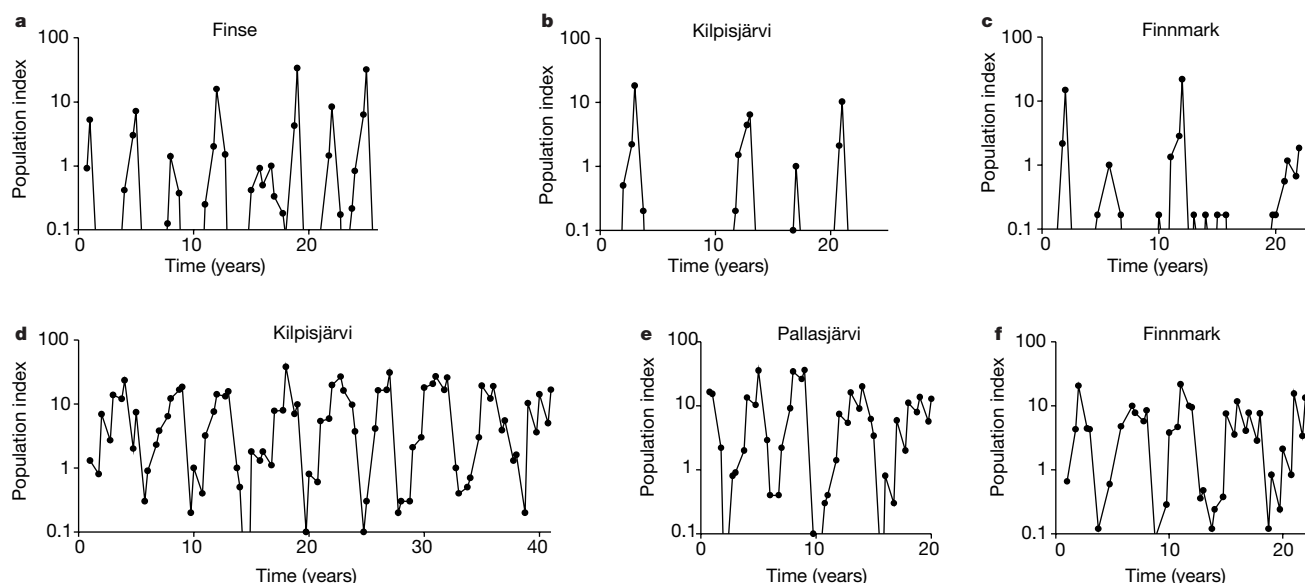


Figure 2 Population dynamics of lemmings at Finse (a), Kilpisjärvi (b) and Finnmark (c), and of voles at Kilpisjärvi (d), Pallasjärvi (e) and Finnmark (f).

therefore experience a winter crash. If this threshold is reached in the fall, then peak density would be equal to $N_{\text{threshold}}$. However, if $N_{\text{threshold}}$ is achieved in the spring, then predator density will increase much beyond it before the winter crash. As a result, peak density of predators should be highly variable (in fact, this mechanism, the interaction between seasonality and predator–prey dynamics, is at the root of mathematical chaos that may arise in rodent population models^{11,14,15}). Again, numerical results are consistent with our main hypothesis (Table 1).

Our findings, which suggest different causal mechanisms for lemming versus vole cycles, concur with what we know about the biology of these rodents. Voles are folivores. Their food plants regrow rapidly after defoliation¹⁶ by mobilizing energy stored underground, and this feature of the vole–vegetation interaction is strongly stabilizing¹³. Lemmings, in contrast, are moss-eaters (particularly so during the critical winter period). Mosses regrow slowly after being depleted by herbivores, and the inherent time lag in this process is highly destabilizing¹³. Furthermore, unlike in more productive vole habitats, lemmings tend to deplete their forage in high arctic and alpine habitats before their predators can reach densities high enough to affect their dynamics. It has been proposed¹⁷ that the dominant consumer–resource interaction in terrestrial grazing webs shifts from the herbivore–plant interface in low productivity systems (lemmings) to the predator–prey interface in higher productivity locations (voles). This view is strongly supported by observations of the strong impact of lemmings on their food supply during peak years^{18–20}. Furthermore, lemmings are much more mobile during population peaks and collapses than voles²¹. This observation makes sense within the framework of the hypothesis that we advocate here. Because vole crashes are caused by predators, moving is futile, and can only increase the risk of being eaten. Lemmings, in contrast, have to move during peak density because they face a desperate shortage of food resources.

Historically, population ecologists tended to look for ‘universal’ explanations of the population dynamics of small rodents³. As a result, when the view gained ground that dynamics of voles *Microtus agrestis* in Fennoscandia are explained by their interactions with specialist and generalist predators^{11,22–25}, it could be argued that lemming cycles are also driven by their predators. Predation hypothesis, however, makes a specific prediction about the topology of lemming cycles, which is at variance with the empirically observed patterns. Thus, our analytical results indicate that popula-

tion oscillations in lemmings and voles may be driven by very different ecological mechanisms. □

Methods

Sources of data

Voles: Kilpisjärvi and Pallasjärvi²⁶, Finnmark²⁷. Lemmings: Finse^{5,28}, Kilpisjärvi (H.H., unpublished data), Finnmark²⁷. Finse and Finnmark lemming data were collected in alpine highlands (the optimal lemming habitat), but Kilpisjärvi data were collected in a birch forest (surrounded by large alpine areas). Birch forest probably constitutes a sink habitat for lemmings, and therefore primarily reflects the lemming dynamics in the surrounding alpine habitat, as indicated by comparing this dataset to once-a-year trapping results from the alpine zone^{29,30}.

Statistical analyses

Skewness. Our theoretical prediction is that log-transformed prey numbers should be negatively skewed, whereas consumers should not be skewed. However, if there are many zero-density observations, and these are assigned to some arbitrary low positive number (which is the standard practice in the analysis of population time-series data; see broken line in Fig. 1), then we should observe an (artificial) positive skewness. The first column in Table 1 (‘log-transformed’) reports the skewness of log-transformed data where zeros are replaced with 0.01; the second column reports the skewness of the frequency distribution of data with zeros omitted; the third column reports the skewness of non-transformed data (where no special handling of zeros is needed). In all cases, vole skewness is significantly less than that for lemmings, as judged by a *t*-test with 4 degrees of freedom. **Pre-peak rate of increase.** This was calculated according to $r_{\text{pre-peak}} = \ln N_1 - \ln N_2$, where N_1 is the density during the peak year, and N_2 is the density of the year before. We defined peak as the year with highest geometric mean of spring and fall density (but other reasonable definitions of peak did not affect the main result). The column ‘*n*’ reports the number of peaks on which $r_{\text{pre-peak}}$ means are based.

Variability in peak density. Using the same definition of peaks as above, we calculated the coefficient of variation (standard deviation divided by the mean) of peak densities for each data series.

Received 15 November 1999; 7 April 2000.

- Elton, C. S. Periodic fluctuations in the number of animals: their causes and effects. *Br. J. Exp. Biol.* **2**, 119–163 (1924).
- Stenseth, N. C. & Ims, R. A. in *The Biology of Lemmings* (eds Stenseth, N. C. & Ims, R. A.) 3–34 (Linnean Society, London, 1993).
- Chitty, D. *Do Lemmings Commit Suicide? Beautiful Hypothesis and Ugly Facts* (Oxford Univ. Press, New York, 1996).
- Stenseth, N. C. & Ims, R. A. in *The Biology of Lemmings* (eds Stenseth, N. C. & Ims, R. A.) 61–96 (Linnean Society, London, 1993).
- Stenseth, N. C., Chan, K. S., Framstad, E. & Tong, H. Phase- and density-dependent population dynamics in Norwegian lemmings: interaction between deterministic and stochastic processes. *Proc. R. Soc. Lond. B* **265**, 1957–1968 (1998).
- Rosenzweig, M. L. & MacArthur, R. H. Graphical representation and stability conditions of predator–prey interaction. *Am. Nat.* **97**, 209–223 (1963).
- May, R. M. in *Theoretical Ecology: Principles and Applications* (ed. May, R. M.) 78–104 (Sinauer Associates, Sunderland, Massachusetts, 1981).

8. Bazykin, A. D., Berezovskaya, F. S., Denisov, G. A. & Kuznetsov, Y. A. The influence of predator saturation effect and competition among predators on predator-prey system dynamics. *Ecol. Mod.* **14**, 39–57 (1981).
9. Beddington, J. R., Free, C. A. & Lawton, J. H. Dynamic complexity in predator–prey model framed in simple difference equations. *Nature* **225**, 58–60 (1976).
10. Oksanen, L. Exploitation ecosystems in seasonal environments. *Oikos* **57**, 14–24 (1990).
11. Hanski, I., Turchin, P., Korpimäki, E. & Henttonen, H. Population oscillations of boreal rodents: regulation by mustelid predators leads to chaos. *Nature* **364**, 232–235 (1993).
12. Turchin, P. & Hanski, I. An empirically-based model for the latitudinal gradient in vole population dynamics. *Am. Nat.* **149**, 842–874 (1997).
13. Turchin, P. & Batzli, G. Availability of food and the population dynamics of arvicoline rodents. *Ecology* (in the press).
14. Oksanen, L. & Oksanen, T. Long-term microtine dynamics in north fennoscandian tundra—the vole cycle and the lemming chaos. *Ecography* **15**, 226–236 (1992).
15. Turchin, P. Chaos and stability in rodent population dynamics: evidence from nonlinear time-series analysis. *Oikos* **68**, 167–172 (1993).
16. Ostfeld, R. S., Canham, C. D. & Pugh, S. R. Intrinsic density-dependent regulation of vole populations. *Nature* **366**, 259–261 (1993).
17. Oksanen, L., Fretwell, S. D., Arruda, J. & Niemela, P. Exploitation ecosystems in gradients of primary productivity. *Am. Nat.* **118**, 240–261 (1981).
18. Batzli, G. O., White, R. G., MacLean, S. F., Pitelka, F. A. & Collier, B. D. In *An arctic ecosystem: the coastal tundra at Barrow, Alaska*. (eds Brown, J., Miller, P. C., Tieszen, L. L. & Bunnell, F. L.) 335–410 (Hutchinson and Ross, Stroudsburg, PA, 1980).
19. Chernyavsky, F. B. & Tkachev, A. V. *Population Cycles of Lemmings in the Arctic: Ecological and Endocrine Aspects* (in Russian) (Nauka, Moscow, Russia, 1982).
20. Moen, J., Lundberg, P. A. & Oksanen, L. Lemming grazing on snowbed vegetation during a population peak. *Arctic Alpine Res.* **25**, 130–135 (1993).
21. Henttonen, H. & Kaikusalo, A. in *The Biology of Lemmings* (eds Stenseth, N. C. & Ims, R. A.) 157–186 (Academic, London, 1993).
22. Hansson, L. & Henttonen, H. Gradients in density variations of small rodents: the importance of latitude and snow cover. *Oecologia* **67**, 394–402 (1985).
23. Hanski, I., Hansson, L. & Henttonen, H. Specialist predators, generalist predators, and the microtine rodent cycle. *J. Anim. Ecol.* **60**, 353–367 (1991).
24. Hanski, I., Henttonen, H., Korpimäki, E., Oksanen, L. & Turchin, P. Small rodent dynamics and predation. *Ecology* (in the press).
25. Korpimäki, E. & Norrdahl, K. Experimental reduction of predators reverses the crash phase as small rodent cycles. *Ecology* **79**, 2448–2455 (1998).
26. Henttonen, H. & Hanski, I. in *Chaos in Real Data* (eds Perry, J. N., Smith, R. H., Woivod, I. P. & Morse, D.) 73–96 (Kluwer Academic, Dordrecht, Netherlands, 2000).
27. Ekerholm, P., Oksanen, L. & Oksanen, T. Long-term dynamics of voles and lemmings in low arctic tundra and above the willow limit as a test of theories on trophic interactions. *Ecography* (in the press).
28. Framstad, E., Stenseth, N. C., Bjørnstad, O. N. & Falck, W. Limit cycles in Norwegian lemmings: tensions between phase-dependence and density dependence. *Proc. R. Soc. Lond. B* **264**, 31–38 (1997).
29. Henttonen, H., Kaikusalo, A., Tast, J. & Viitala, J. Interspecific competition between small rodents in subarctic and boreal ecosystems. *Oikos* **29**, 581–590 (1977).
30. Virtanen, R., Henttonen, H. & Laine, K. M. Lemming grazing and the structure of snowbed plant community—a long-term experiment at Kilpisjärvi, in Finnish Lapland. *Oikos* **79**, 155–166 (1997).

Acknowledgements

We thank I. Hanski and N. C. Stenseth for valuable discussion and comments on the manuscript.

Correspondence and requests for materials should be addressed to P.T. (e-mail: Peter.Turchin@UConn.edu).

Highly fecund mothers sacrifice offspring survival to maximize fitness

Sigurd Einum*† & Ian A. Fleming*

* Norwegian Institute for Nature Research, Tungasletta 2, N-7485 Trondheim, Norway

† Department of Zoology, Norwegian University of Science and Technology, N-7034 Trondheim, Norway

Why do highly fecund organisms apparently sacrifice offspring size for increased numbers when offspring survival generally increases with size^{1–3}? The theoretical tools for understanding this evolutionary trade-off between number and size of offspring have developed over the past 25 years^{1,4–10}; however, the absence of data on the relation between offspring size and fitness in highly fecund species, which would control for potentially confounding variables, has caused such models to remain largely

hypothetical^{11,12}. Here we manipulate egg size, controlling for maternal trait interactions, and determine the causal consequences of offspring size in a wild population of Atlantic salmon. The joint effect of egg size on egg number and offspring survival resulted in stabilizing phenotypic selection for an optimal size. The optimal egg size differed only marginally from the mean value observed in the population, suggesting that it had evolved mainly in response to selection on maternal rather than offspring fitness. We conclude that maximization of maternal fitness by sacrificing offspring survival may well be a general phenomenon among highly fecund organisms.

One of the most intensely studied maternal traits is egg size, owing to its direct consequences for offspring fitness^{1,3}. Individual mothers are forced to trade off offspring quantity against quality during reproduction^{13,14}. Much of the theoretical work on this question of reproductive allocation has focused around the seminal paper of Smith and Fretwell⁴. Because of the trade-off between egg size and number, their model predicts the evolution of an optimal egg size that maximizes maternal fitness (product of number of offspring and their fitness), with mothers having the upper hand in this parent–offspring conflict. This potentially explains the existence of species experiencing massive mortality during juvenile stages, because offspring survival rates that maximize maternal fitness may be low compared with those maximizing offspring fitness. When considering the impact that the Smith–Fretwell model has had on aspects of life-history theory^{1,5–10}, it is worrying that empirical support from highly fecund species is lacking. The general validity of the Smith–Fretwell model has previously been questioned on the basis of data from organisms producing few offspring¹², and it has been criticized for being too simplistic¹¹.

We therefore undertook an experimental field study using Atlantic salmon to test empirically whether maternal or offspring fitness is maximized in highly fecund species (egg numbers range from 1,791 to 18,847)¹⁵. Egg size was manipulated by rearing females to adulthood in captivity. This procedure resulted in some females producing highly variable egg sizes (mean coefficient of variation = 18.5%) relative to that found in the wild (4.0%; S.E. and I.A.F., unpublished data). A sample of small and large eggs from each of eight females was fertilized by one male, producing a total of eight pairs of full-sib groups. At the eyed stage, equal numbers of small and large eggs within a pair were buried in separate artificial gravel nests³, producing a total of 16 such nests. Small eggs were on average 31.4% lighter than their large siblings ((mean $\pm$ s.d.) small: 85.4 $\pm$ 15.1 mg; large: 125.4 $\pm$ 13.4 mg; t_7 = 6.06, P = 0.001, paired samples t -test). The size separation within family groups was responsible for 69.2% of the total variation in egg size within the population, randomizing the effects of other maternal or genetic traits potentially correlated with egg size. Thus, this design allowed us to test for a causal relationship between egg size and offspring fitness^{16,17}. It excluded the possible effects of interactions with other maternal traits that may cause the optimum egg size to vary among females within natural populations^{1,6,18}. Juveniles emerging from the nests were released in a natural stream, and were sampled 28 and 107 d after median emergence to assess their success.

Table 1 Selection on egg size for Atlantic salmon

	β	β'	P	γ	γ'	P
Offspring survival	19.65 (5.05)	0.488 (0.126)	0.001	–573.8 (343.5)	–14.30 (8.54)	0.069
Maternal RS	14.43 (5.85)	0.358 (0.145)	0.009	–723.7 (305.5)	–18.02 (7.59)	0.023

Selection gradients relate relative offspring survival and maternal reproductive success (RS, proportion offspring survival $\times$ potential maternal fecundity) to egg size. Directional selection gradients (unstandardized, β ; standardized, β') are estimated from linear regression coefficients, and nonlinear (stabilizing) selection gradients (unstandardized, γ ; standardized, γ') are estimated from regression coefficients of squared deviations from the mean²¹. Standard errors are indicated in parentheses. Residuals from regressions between the traits and log-transformed fitness measures proved to be normally distributed, allowing parametric tests of significance²¹.

Egg size affected both timing of, and size at emergence from, gravel nests. Juveniles from small eggs emerged from the nests on average 3 d earlier than their siblings from large eggs ($t_7 = 7.32$, $P < 0.001$, paired t -test). At this point, juveniles from small eggs weighed 33.1% less than their siblings from large eggs ((mean $\pm$ s.d.) small: 115.9 ± 21.3 mg; large: 174.6 ± 19.1 mg; $t_7 = 6.18$, $P < 0.001$). At recapture, 28 and 107 d after median date of emergence from nests, there was still a strong positive effect of egg size on body size (first recapture: body weight = $-0.02 \text{ g} + 4.16 \times \text{egg weight}$, $r^2 = 0.81$, $P < 0.001$; second recapture: body weight = $0.82 \text{ g} + 8.02 \times \text{egg weight}$, $r^2 = 0.43$, $P = 0.011$). Thus, egg size had strong effects on traits closely linked with fitness^{2,3}.

During the first period (0–28 d after median emergence) juveniles from large eggs experienced significantly higher survival, measured as proportion recaptured, than their siblings from small eggs ((mean proportion recaptured $\pm$ s.d.); small: 0.048 ± 0.046 ; large: 0.145 ± 0.043 ; $t_7 = 5.46$, $P = 0.001$, paired test). The same pattern also appeared when considering the effect of mortality within the stream alone, that is, removing those caught in the drift samples from the release numbers (small: 0.059 ± 0.054 ; large: 0.184 ± 0.036 ; $t_7 = 5.39$, $P = 0.001$), and when considering fish caught in drift samples as survivors (small: 0.252 ± 0.074 ; large: 0.363 ± 0.105 ; $t_7 = 3.22$, $P = 0.015$). Consistent with other results², there was no additional effect of egg size on survival during the second period (28–107 d after median emergence, $n = 14$, $r = -0.12$,

$P = 0.690$). Thus, maternal traits appear to be most intensely shaped during early juvenile stages, and less so later on when expression of offspring genes may become more important for their success. The overall relation between egg size and proportion recaptured was best described by asymptotic regressions (Fig. 1).

Estimated selection gradients demonstrate strong directional selection towards larger egg size (Table 1). Furthermore, a significant negative quadratic term for maternal reproductive success suggests that selection acted directly to reduce the variance in egg size through stabilizing selection. We regard these results as a confirmation of the Smith–Fretwell model, legitimizing the use of its assumptions and predictions in development of other life-history models^{5–10}. Moreover, there was a close fit between optimal and observed egg size (Fig. 1). This indicates that mean egg size in Atlantic salmon has evolved as a response to selection on maternal fitness.

When considering the strong phenotypic selection on egg size, one might expect this trait to be relatively invariable. Thus, large intrapopulation variation in egg size, commonly observed in a wide range of organisms¹ as well as in the study population (Fig. 1), raises an interesting evolutionary problem. Functional constraints on egg size caused by morphological features in certain taxa (for example reptiles) have been suggested to cause such variation^{12,14}. Furthermore, several hypotheses explain intrapopulation variation in egg size as adaptive phenotypic plasticity caused by interactions between egg size and correlated maternal traits on offspring fitness^{1,6,18}. Although the problem of intrapopulation variation can not be solved by traditional optimality models¹¹, this does not imply that these models are incomplete as a means of predicting mean optimal egg size, as successfully shown here. Rather, further development of models and empirical work on evolution of reproductive traits will probably benefit from applying and extending the basic assumptions and predictions of these models. □

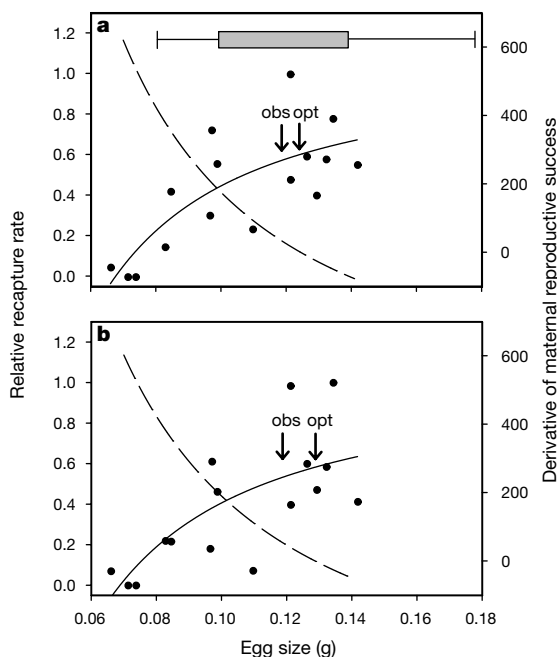


Figure 1 Relation between egg size and relative recapture rate (scaled to a maximum of 1) of juvenile Atlantic salmon 28 (a) and 107 (b) days after median emergence from nests. Each data point represents mean values for one nest. The asymptotic functions (solid curves) relating offspring recapture rates to egg size (m) are of the form $F(m) = 1 - (m_{\min}/m)^a$, where $m_{\min}$ is the minimum viable egg size and a is a constant⁸. Dashed lines represent the derivative of the function relating maternal reproductive success to egg size (reproductive success = $F(m)M/m$) given a total egg mass (M) of 450 g (mean of the native population¹⁸). The derivative equals zero for the egg size that maximizes maternal reproductive success⁶ (opt), assuming variation in total egg mass among females does not affect offspring survival (for example, the intensity of offspring sibling competition is unrelated to female fecundity⁹). Observed mean (obs), and standard deviation and range (box plot) of egg size of the native population¹⁸ (adjusted for increase in size owing to absorption of water, $16.4 \pm 3.4\%$, $n = 23$ females; S.E. and I.A.F., unpublished data) are indicated. $m_{\min} = 0.0676$ g, $a = 1.5066$, $r^2 = 0.59$, $P < 0.001$ (a); $m_{\min} = 0.0688$ g, $a = 1.3950$, $r^2 = 0.50$, $P = 0.002$ (b).

Methods

The adults used originated from gametes obtained from wild fish of the study population. The resulting embryos were incubated in the hatchery and from the onset of exogenous feeding, the juveniles were reared in 2-m² indoor, flow tanks for three months. Thereafter, they were reared in 8-m² freshwater, outdoor tanks (density during maturation 350 fish (1 kg) per tank) under a natural light regime and fed commercial food *ad lib*. The resulting egg size variation was within the range observed in females in nature¹⁵. Such variability is probably related to the position of the egg relative to blood vessels in the female ovary during development¹⁹. The relative chemical composition (that is, percentage of water, protein and lipids) and weight-specific energy content of different sized eggs from females reared in this way has previously been found not to differ³. Furthermore, under favourable hatchery conditions offspring from small eggs performed as well as siblings from large eggs³. Thus, the obtained variability in egg size appears to be independent of genetic, or other characteristics of the oocytes themselves.

Juveniles emerging from the nests were anaesthetized, weighed (± 0.1 mg), and group marked using alcan blue dye and adipose fin clips. They were then held for 24 h in an enclosure submerged in the stream Ålabekk (see ref. 20) to ensure recovery before being released. A total of 1,401 emerging juveniles were released, all at the same location. Losses from the population through migration were controlled by an inaccessible waterfall above and drift nets below the 140-m long experimental area. Body size and survival was assessed by sampling 28 d after median emergence, during which Ålabekk was electrofished five times. Juveniles were then re-marked and re-released, before being sampled again using the same procedure 107 d after median emergence. There were no indications that larger fish within the size range studied were easier to catch, as there were no significant differences in the size of fish captured on different days ((mean weight $\pm$ s.d.) first period: day 1, 0.42 ± 0.14 , day 2, 0.48 ± 0.10 , day 3, 0.47 ± 0.07 , day 4, 0.46 ± 0.08 , $F_{3,132} = 1.55$, $P = 0.204$; second period: day 1, 1.80 ± 0.55 , day 2, 1.64 ± 0.47 , day 3, 2.09 ± 0.00 , $F_{2,82} = 0.86$, $P = 0.427$; see also ref. 2).

Received 8 December 1999; accepted 15 March 2000.

- Roff, D. A. *The Evolution of Life Histories: Theory and Analysis* (Chapman & Hall, New York, 1992).
- Einum, S. & Fleming, I. A. Selection against late emergence and small offspring in Atlantic salmon (*Salmo salar*). *Evolution* **54**, 628–639 (2000).
- Einum, S. & Fleming, I. A. Maternal effects of egg size in brown trout (*Salmo trutta*): norms of reaction to environmental quality. *Proc. R. Soc. Lond. B* **266**, 2095–2100 (1999).
- Smith, C. C. & Fretwell, S. D. The optimal balance between size and number of offspring. *Am. Nat.* **108**, 499–506 (1974).
- Brockelman, W. Y. Competition, the fitness of offspring, and optimal clutch size. *Am. Nat.* **109**, 677–699 (1975).

6. Parker, G. A. & Begon, M. Optimal egg size and clutch size: effects of environment and maternal phenotype. *Am. Nat.* **128**, 573–592 (1986).
7. Lloyd, D. C. Selection of offspring size at independence and other size-versus-number strategies. *Am. Nat.* **129**, 800–817 (1987).
8. McGinley, M. A., Temme, D. H. & Geber, M. A. Parental investment in offspring in variable environments: theoretical and empirical considerations. *Am. Nat.* **130**, 370–398 (1987).
9. Sargent, R. C., Taylor, P. D. & Gross, M. R. Parental care and the evolution of egg size in fishes. *Am. Nat.* **129**, 32–46 (1987).
10. Geritz, S. A. H. Evolutionary stable seed polymorphism and small-scale spatial variation in seedling density. *Am. Nat.* **146**, 685–707 (1995).
11. Bernardo, J. The particular maternal effect of propagule size, especially egg size: patterns, models, quality of evidence and interpretations. *Am. Zool.* **36**, 216–236 (1996).
12. Sinervo, B., Doughty, P., Huey, R. B. & Zamudio, K. Allometric engineering: a causal analysis of natural selection on offspring size. *Science* **258**, 1927–1930 (1992).
13. Elgar, M. A. Evolutionary compromise between a few large and many small eggs: comparative evidence in teleost fish. *Oikos* **59**, 283–287 (1990).
14. Sinervo, B. & Licht, P. Proximate constraints on the evolution of egg size, number, and total clutch mass in lizards. *Science* **252**, 1300–1302 (1991).
15. Fleming, I. A. Reproductive strategies of Atlantic salmon: ecology and evolution. *Rev. Fish Biol. Fish.* **6**, 379–416 (1996).
16. Wade, M. J. & Kalisz, S. The causes of natural selection. *Evolution* **44**, 1947–1955 (1990).
17. Sinervo, B. & Svensson, E. Mechanistic and selective causes of life history trade-offs and plasticity. *Oikos* **83**, 432–442 (1998).
18. Jonsson, N., Jonsson, B. & Fleming, I. A. Does early growth cause a phenotypically plastic response in egg production of Atlantic salmon? *Funct. Ecol.* **10**, 89–96 (1996).
19. Kamler, E. *Early Life History of Fish: an Energetics Approach* (Chapman & Hall, London, 1992).
20. Einum, S. & Fleming, I. A. Genetic divergence and interactions in the wild among native, farmed and hybrid Atlantic salmon. *J. Fish Biol.* **50**, 634–651 (1997).
21. Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226 (1983).

Acknowledgements

We thank J. G. Backer, K. Bergersen, T. Husebø and the rest of the staff at the NINA Research Station, Ims, for assistance with the experiments. We also thank J. D. Armstrong, P. Fiske, T. Forseth, A. P. Hendry, K. Hindar, B. Jonsson and O. Ugedal for comments on the manuscript. Financial support was provided by a PhD scholarship from the Norwegian Research Council to S.E. and research funding from the Norwegian Institute for Nature Research to I.A.F.

Correspondence and requests for materials should be addressed to S.E. (e-mail: sigurd.einum@ninatrd.ninaniku.no).

Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task

Mark Laubach, Johan Wessberg & Miguel A. L. Nicolelis

Department of Neurobiology, Duke University Medical Centre, Durham, North Carolina 27710, USA

When an animal learns to make movements in response to different stimuli, changes in activity in the motor cortex seem to accompany and underlie this learning^{1–6}. The precise nature of modifications in cortical motor areas during the initial stages of motor learning, however, is largely unknown. Here we address this issue by chronically recording from neuronal ensembles located in the rat motor cortex, throughout the period required for rats to learn a reaction-time task. Motor learning was demonstrated by a decrease in the variance of the rats' reaction times and an increase in the time the animals were able to wait for a trigger stimulus. These behavioural changes were correlated with a significant increase in our ability to predict the correct or incorrect outcome of single trials based on three measures of neuronal ensemble activity: average firing rate, temporal patterns of firing, and correlated firing. This increase in prediction indicates that an association between sensory cues and movement

emerged in the motor cortex as the task was learned. Such modifications in cortical ensemble activity may be critical for the initial learning of motor tasks.

Traditional theories of cortical function propose that information in the central nervous system is represented primarily by the firing rate of single neurons^{7,8}. Learning new motor contingencies should therefore primarily involve changes in the average firing of neurons located in sensorimotor cortical areas. Behaviourally relevant information can also be represented in temporal patterns of neuronal firing (on a ms scale) and correlated neuronal activity^{9–15}. It is not known whether these other schemes of neuronal representation are significant in the physiological modifications associated with the initial learning of a motor task. To address this issue, we compared the potential contribution of three neuronal coding schemes (average firing rate, temporal patterns of firing, and correlated activity) for predicting the behavioural outcome of single trials throughout the period required for rats to learn a reaction-time task. We also compared concurrent neural ensemble and electromyography (EMG) recordings to establish that any increase in single trial prediction by ensemble activity was not due to differences in movements on the correct and error trials.

Rats were trained to perform a reaction-time task in which they were required to hold down a response lever throughout a variable interval (the foreperiod, 400–800 ms) and to release the lever in response to vibrotactile or auditory trigger stimuli with a short (< 1 s) reaction time (Fig. 1a). Correct trials occurred when the rats sustained the lever press until the presentation of the trigger stimuli and released the lever with a short reaction time. Error trials occurred when the rats released the lever before the trigger stimulus. During the first of week of training, the rats exhibited significant increases in the time they were able to wait for the trigger stimulus, from 422.3 ± 30.9 ms (day 1) to 577.8 ± 24.8 ms (day 7; repeated-measures analysis of variants (ANOVA): $P < 10^{-4}$; Fig. 1b). Although the rats' reaction times during these sessions did not change (day 1: 320.8 ± 41.5 ms; day 7: 271.6 ± 32.9 ms;

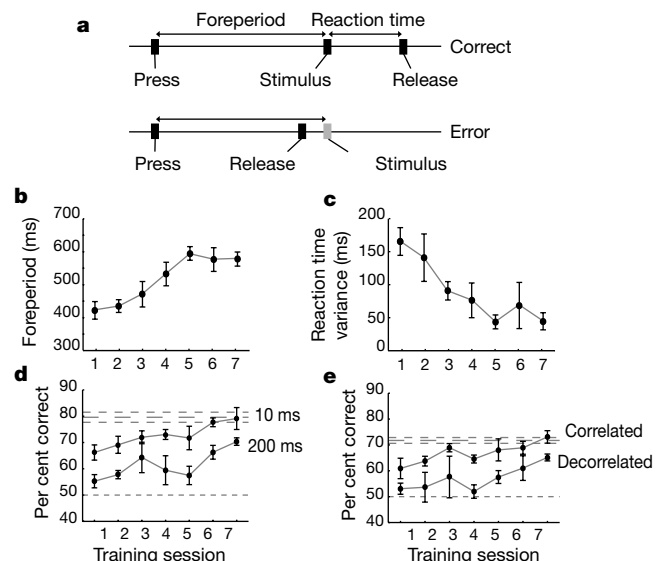


Figure 1 Improvements in behavioural performance paralleled increases in the neuronal predictions of trial outcomes. **a–c**, As the rats learned the task (**a**), they sustained lever presses over longer foreperiods (**b**) and exhibited decreased variance in their reaction times (**c**). **d**, Predictions of trial outcomes based on neuronal activity increased with training and, by day 6–7, achieved levels equal to those in fully trained rats (upper lines, mean $\pm$ s.e.m.). Chance discrimination was 50% (lower dashed line). Temporal patterns of firing (from smoothing over 10-ms intervals) predicted trial outcomes better than average firing rates (combined over 200 ms). Likewise, predictions based on correlated firing were significantly better than those based on decorrelated ensemble activity.

repeated-measures ANOVA: $P > 0.05$), the variance of the rats' reaction times decreased significantly (day 1: 165 ± 24.2 ms; day 7: 44.5 ± 6.6 ms; repeated-measures ANOVA: $P < 0.02$; Fig. 1c). After these training sessions, our animals performed the task as well as animals in previous studies of reaction-time behaviour in rats, that is, better than 70% correct trials¹⁶.

Chronic and simultaneous recordings were obtained from neurons (25.4 ± 1.3 neurons per session, mean $\pm$ s.e.m., range 14 to 32 neurons, total 711) in the forelimb representation of two motor areas of the cerebral cortex (medial and lateral agranular cortical areas, mAG and lAG) of six rats. To study the effect of learning on neuronal activity, spike trains in the 200-ms interval before lever release were compared for correct and error trials. The capability of three different types of neuronal coding schemes to predict the outcomes of single trials as correct or error was quantified using artificial neural network (ANN) methods^{17,18}. The ANNs were trained with neuronal ensemble responses from subsets of trials and tested with the remaining trials (Fig. 2) for each training session throughout the learning period. The degree to which confusion matrices from the ANNs differed from chance was then evaluated using information theory¹⁹ and was expressed as bits of information.

The majority of cortical neurons (477, 67.1%) were active when the rats pressed and released the lever (Fig. 2). However, over the seven training sessions, the responses of only nine of these neurons provided more than 0.1 bits of information about trial outcome. That is, 98.1% of the single neurons alone did not predict trial outcome accurately (overall average of $59.3 \pm 0.3\%$ correct, 0.015 ± 0.001 bits for all training sessions). The average amount of information represented by the single neurons was 0.011 ± 0.001 bits on the first day of training and 0.024 ± 0.004 bits on the seventh day of training. Permutation tests²⁰ showed that these information values are not statistically significant and can be generated by chance alone.

By contrast, predictions of trial outcome based on the average firing rate (using a single 200-ms bin) of the same cortical neurons analysed together as ensembles improved significantly in the first seven days of training, from $55.3 \pm 2.5\%$ (0.013 ± 0.004 bits, day 1) to $70.3 \pm 1.4\%$ (0.073 ± 0.017 bits, day 7, $P < 0.01$, repeated measures ANOVA, Fig. 1d). While the amount of information represented by average ensemble firing rate on the seventh day of training was small, permutation tests showed that it was highly significant ($P \approx 0.30$, day 1; $P = 0.02$, day 7). This improvement in trial prediction over learning occurred in the absence of any significant change in overall firing rates for the neuronal ensembles (20.6 ± 1.6 Hz for all sessions, Kruskal–Wallis test: $P = 0.573$).

Next, we observed that single trial classification improved significantly when neuronal spike trains were smoothed over 10-ms intervals (Fig. 3a), from $66.2 \pm 2.9\%$ (0.087 ± 0.020 bits, day 1) to $79.1 \pm 4.2\%$ of the trials correct (0.227 ± 0.060 bits, day 7). This corresponded to more than a 200% increase in trial prediction based on information theory ($P < 0.01$, repeated measures ANOVA, Fig. 1d). Predictions based on fine temporal patterns of firing (10-ms intervals) on the seventh day of training yielded three times more information than average firing rates (200-ms interval) and were better than those based on average firing rates for all days of training (ANOVA: $P < 10^{-4}$). In fully trained animals ($n = 6$), the outcomes of single trials were predicted equally well if spikes were smoothed over 5-ms intervals ($79.6 \pm 1.9\%$ correct, 0.207 ± 0.032 bits) or 10 ms ($77.3 \pm 1.4\%$, 0.178 ± 0.025 bits). However, predictions were significantly worse (ANOVA: $P < 0.005$) when spikes were smoothed over 20-ms intervals ($71.8 \pm 0.7\%$ correct, 0.105 ± 0.008 bits) or combined over 200 ms ($63.6 \pm 1.9\%$ correct, 0.038 ± 0.011 bits). These results indicate that fine temporal integration (10 ms or less) of firing across an ensemble of cortical neurons may encode the contingency between trigger stimuli and movement execution.

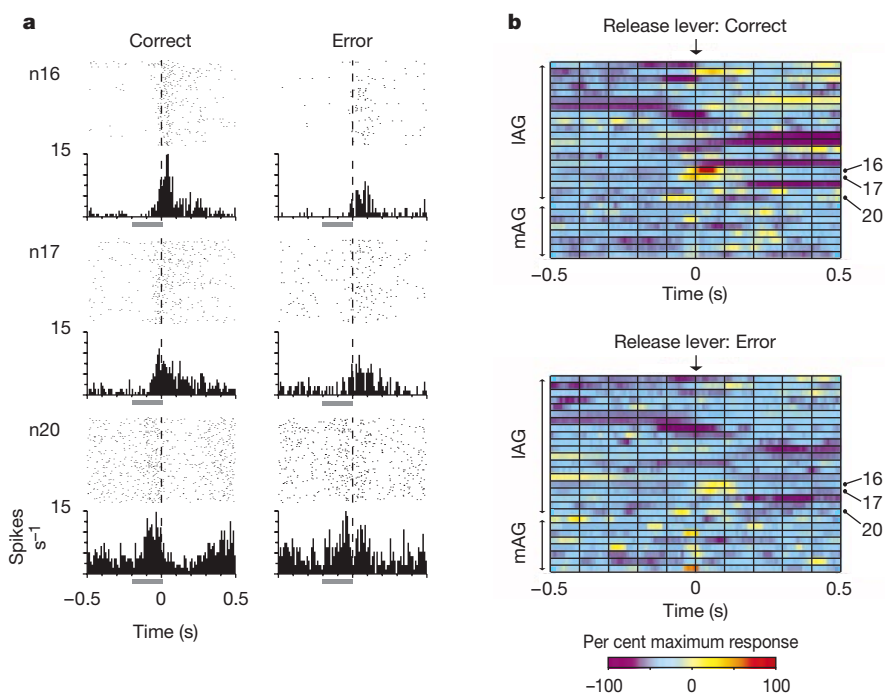


Figure 2 Neuronal ensemble responses associated with correct and error trials. **a**, Single neurons (n) in the lateral (lAG) and medial (mAG) agranular areas of motor cortex fired differently on correct and error trials (bin, 10 ms). The bars below each plot show the 200-ms period analysed in this study. **b**, Different temporal modulations of firing on correct and

error trials occurred over the neuronal ensembles around lever release (time = 0). Each horizontal line represents the firing of a neuron over time (in seconds). Neuronal firing was normalized to be -100 (purple) to $+100\%$ (red) of the maximum firing rate in the ensemble (bin, 10 ms; responses smoothed with a 5-ms moving average filter).

We also measured how well correlated firing predicted single trial outcomes over the course of learning. Independent component analysis (ICA; refs 21, 22) was used to quantify zero-lag correlations between groups of neurons (Fig. 3b). Prediction of single trial outcome based on the independent components improved significantly over learning, from $60.9 \pm 4.0\%$ (0.049 ± 0.022 bits on day 1) to $73.1 \pm 2.4\%$ (0.167 ± 0.044 bits on day 7, repeated measures ANOVA: $P < 0.01$, Fig. 1e), and provided more than twice the information obtained from average firing rate. The amount of covariance accounted for by the independent components did not change over the training period (day 1: $42.6 \pm 2.3\%$; day 7: $45.0 \pm 4.5\%$; repeated measures ANOVA: $P = 0.85$). This result indicates that although the degree of correlated activity within the ensembles did not change, the information about trial outcome contained in the correlated activity increased with learning. The distributions of the independent components over the neuronal ensembles revealed patterns of correlated firing by subsets of the neurons that were not equally distributed across the motor cortex (Fig. 3b).

To demonstrate further that the information accounted for by ICA was due to correlated firing, a 'spike-shifting' procedure was used to reduce correlations across the neuronal ensemble without altering the firing rates of the individual neurons. This procedure resulted in a significantly lower percentage of correctly classified trials by the ANNs throughout learning ($n = 4$, $P < 0.001$, Fig. 1e). These values of information were not significantly different from those obtained with the average firing rates of the same ensembles (Student's t -test: $P > 0.05$). Thus, correlated firing, as detected by ICA, proved to be a potential means of encoding information in addition to that available in average neuronal ensemble firing rates.

One possible confound for this study was that the occurrence of sensory stimuli on correct but not error trials could have been the

basis of the discriminations. However, several results discount this possibility. First, very few neurons exhibited any response to the vibrotactile and none to the auditory stimuli. Second, in the rare occasions when a tactile response was observed, the latency and duration of the responses were brief. As our data analysis interval was 200 ms before movement onset, and the average reaction time was 271.6 ± 32.9 ms (day 7), these short-lasting tactile responses did not have a major impact on our analyses. Also, the time interval accounting for most of the single trial prediction (Figs 2 and 3) occurred 100 ms before movement onset. Thus, it seems unlikely that sensory responses of the cortical neurons accounted for the improvement in trial prediction observed over learning.

Another potential confound is that the cortical neuronal patterns that produce predictions of single trials may have reflected differences in motor output between correct and error trials. Several types of control data imply that this is not probable. First, video recordings of the animals showed no difference in overt behaviour between the types of trials. Second, there was no statistical difference in lever trajectories between correct and error trials. Third, EMG activity (Fig. 4a, b), recorded from several of the main muscles involved in lever release in two control animals, provided insignificant amounts of information about whether a single trial was correct or not (rat 1, 2 muscles: $62.2 \pm 2.0\%$ correct, 0.008 ± 0.003 bits; rat 2, four muscles: $61.8 \pm 1.1\%$ correct, 0.029 ± 0.007 bits; $P > 0.1$; Fig. 4c). Fourth, in both control animals, single trial classification achieved using groups of muscles was not better than that obtained using the best individual muscle in each animal (that is, there was no additive effect across muscles). Finally, when neuronal ensemble activity in motor cortex was recorded simultaneously with multiple EMGs in the same sessions, we observed that single trial discrimination based on neuronal data continued to be highly significant even though

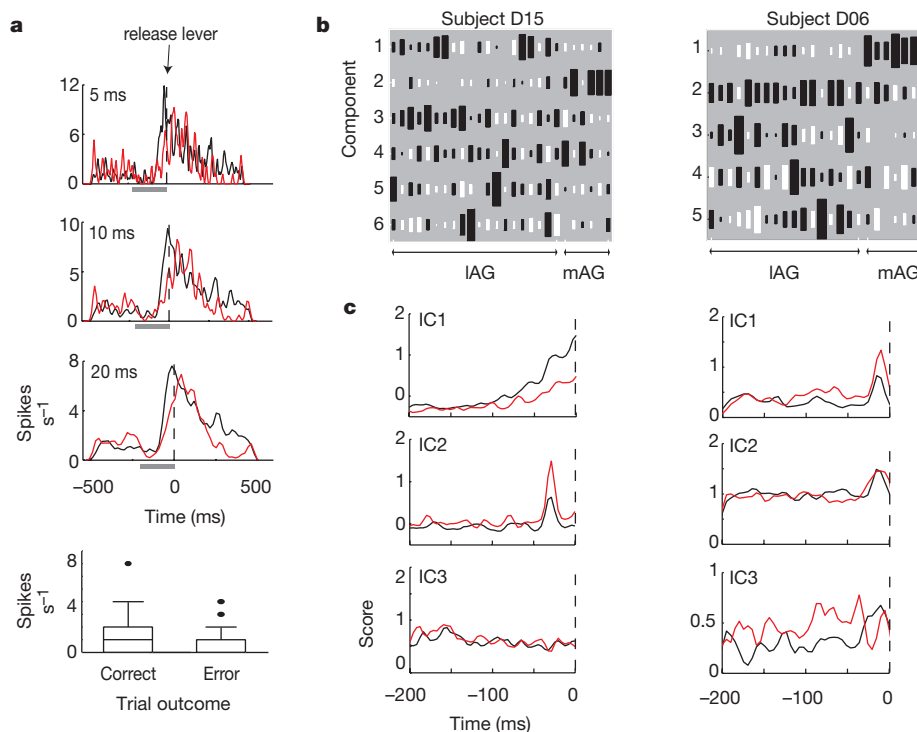


Figure 3 Predictions of trial outcome were based on temporal patterns of firing and correlated activity. **a**, Temporal patterns of neuronal firing (about 5–10 ms) predicted correct (black) and error (red) trials better than average firing rates (box plots: box is interquartile range (IQR), line is median, and dots are extreme values, that is, greater than 1.5 times the IQR), which were not different on correct and error trials. **b,c**, Independent component analysis (ICA) detected correlations between subgroups of neurons (**b**) during

the final 200 ms before movement (data from two fully trained subjects). Neurons that fired together have the same colour boxes, and the size of the boxes reflects the correlation between the neurons. Trial outcomes were predicted significantly above chance due to the occurrence of different patterns of correlated firing (**c**) on correct (black) and error (red) trials.

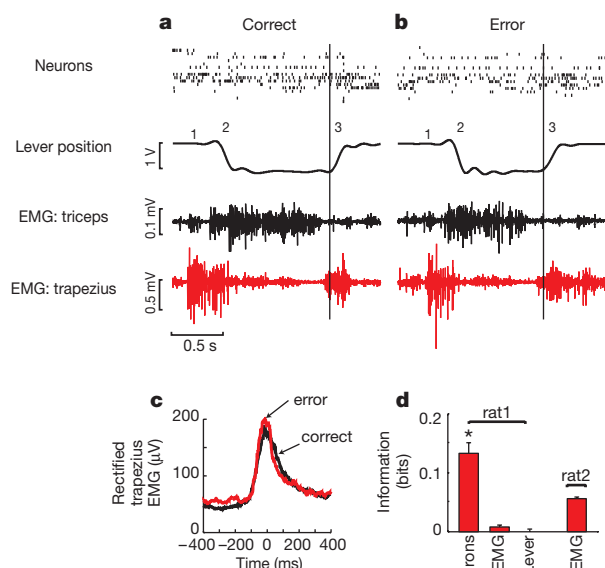


Figure 4 Muscle activations and response kinematics were similar on correct and error trials. **a**, Correct trial. **b**, Error trial. The triceps brachii (black) showed increased EMG activity during lever pressing, while the trapezius (red) showed increased activity as the limb was placed on the lever and during lever release (vertical lines). Behavioural events: 1, placing paw on lever; 2, pressing lever; 3, releasing lever. **c**, Average EMG activations and lever kinematics were equivalent on correct (black) and error (red) trials. **d**, Neuronal ensemble activity recorded in one of these rats provided far more information (asterisk, $P < 0.01$) than the multiple EMG or lever position signals, which did not predict trial outcome better than expected by chance ($P > 0.1$).

EMGs never achieved significance (Fig. 4c). In summary, these control experiments indicate that the improvement in predictions of trial outcomes over the learning period were not because of differences in the movements between correct and error trials.

Our results demonstrate that the ability of neuronal ensembles in the motor cortex to encode sensorimotor contingencies increases during motor learning. We found that the average firing rate, temporal structure and correlated activity of neuronal ensembles in the motor cortex changes during learning to represent new behaviourally relevant information. As reproducible results were obtained in four rats by randomly sampling 20–30 neurons per animal, these changes probably involve widespread territories of the rat motor cortex. Based on these observations, we propose that new functional associations between subgroups of cortical neurons may emerge as animals learn the new sensorimotor contingencies that are required to solve motor tasks. □

Methods

For more detailed methods see Supplementary Information.

Behavioural task

We recorded neuronal responses during learning from four adult, male rats. Neuronal responses were recorded from two additional animals only after they were fully trained, and two more animals were used for control studies using EMG. The rats were moderately deprived of water and trained to press a lever and release it in response to a trigger stimulus for a liquid reward. The foreperiod was initially 250 ms and was incremented by 50–100 ms whenever the rats made three consecutive correct trials. Once the foreperiod was greater than 600 ms and the animal performed the task consistently, the foreperiod varied randomly between 400–800 ms. All events during training and recording were controlled by computer. Behaviour of all rats was monitored by a video system, and by a position sensor attached to the lever in the two rats used for EMG controls.

Electrophysiological recordings

All rats were treated in accordance with NIH guidelines. Procedures for chronically implanting arrays of microwires (NB Labs, Dennison, Texas) and data acquisition, using a Many Neuron Acquisition Program (Plexon, Dallas, Texas), are described elsewhere²³.

After the completion of all experiments, select recording sites were marked with electrolytic lesions before formalin perfusion, and the locations of microwires were reconstructed using histological techniques.

Multiple bipolar patch electrodes (MicroProbe, Potomac, Maryland) were implanted subcutaneously into two rats to record EMG ensemble activity chronically²⁴. In the first rat, we recorded the activity of the triceps brachii (action: elbow extension) and the thoracic portion of trapezius (rotation of the limb at the shoulder). In the second rat, we recorded EMG activity from the anterior portion of deltoideus and pectoralis in addition to trapezius and triceps. EMG signals were amplified (SA Instrumentation, Encinitas, California), filtered at 20–500 Hz and sampled at 1,000 Hz.

Data analysis

Our strategy for statistical analysis of neuronal ensemble data sets is described elsewhere^{18,25,26}. Single trial classification based on average neuronal firing rates was assessed by first summing all spikes for each trial over a 200-ms interval before lever release. These spike counts were then fed into an artificial neural network (ANN) that used the learning vector quantization (LVQ) algorithm¹⁷. Training and testing of artificial neural networks were performed on each behavioural training session in each subject to obtain the numbers depicting single trial classification. Leave-one-out cross-validation²⁷ was used to estimate error rates and the confusion matrices for each data set. The degree to which confusion matrices from the ANNs differed significantly from chance was then evaluated using information theory¹⁹ and permutation tests²⁰, and was expressed as bits of information.

Peri-event histograms (1-ms bins) were constructed for the correct and error trials for the 200-ms interval before lever release. The spike trains were then smoothed over 5-, 10- or 20-ms intervals by low-pass filtering and re-sampling (Fig. 3a). Firing patterns associated with correct and error trials were identified for each neuron using a wavelet-based method²⁸, which essentially identified temporally specific, salient features in the data and thereby reduced the number of variables used to train the ANNs. Classification using the wavelet features was done as described above.

Independent component analysis (ICA) was used to assess how well correlated activity among different groups of cortical neurons predicted single trial outcome²². First, principal component analysis²⁹ was used to reduce the dimensionality of data from 200-ms intervals before lever release. An extended independent component algorithm²¹ was then used to rotate the weights for the largest principal components to make the functions as statistically independent as possible. Scores for the resulting independent components were then computed for each 1-ms interval. These scores were then preprocessed with the wavelet algorithm and analysed with the same artificial neural network methods as described above. To test the notion that the ICA procedure detected correlated firing, the spike trains were decorrelated using a spike-shifting procedure described elsewhere³². Single trial classification based on digitally rectified EMG activity or lever position data was done using the same procedures as for the neuronal spike trains.

Received 8 December 1999; accepted 3 April 2000.

- Mitz, A. R., Godschalk, M. & Wise, S. P. Learning-dependent neuronal activity in the premotor cortex: activity during the acquisition of conditional motor associations. *J. Neurosci.* **11**, 1855–1872 (1991).
- Germain, L. & Lamarre, Y. Neuronal activity in the motor and premotor cortices before and after learning the associations between auditory stimuli and motor responses. *Brain Res.* **611**, 175–179 (1993).
- Chen, L. L. & Wise, S. P. Neuronal activity in the supplementary eye field during acquisition of conditional oculomotor associations. *J. Neurophysiol.* **73**, 1101–1121 (1995).
- Chen, L. L. & Wise, S. P. Supplementary eye field contrasted with the frontal eye field during acquisition of conditional oculomotor associations. *J. Neurosci.* **73**, 1122–1134 (1995).
- Chen, L. L. & Wise, S. P. Evolution of directional preferences in the supplementary field acquisition of conditional oculomotor associations. *J. Neurosci.* **16**, 3067–3081 (1996).
- Chen, L. L. & Wise, S. P. Conditional oculomotor learning: population vectors in the supplementary eye field. *J. Neurophysiol.* **78**, 1166–1169 (1997).
- Georgopoulos, A. P. et al. Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. coding of the direction of movement by a neuronal population. *J. Neurosci.* **8**, 2928–2938 (1988).
- Shadlen, M. N. & Newsome, W. T. Noise, neural code and cortical organization. *Curr. Opin. Neurobiol.* **4**, 569–579 (1994).
- Murthy, V. N. & Fetz, E. E. Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proc. Natl Acad. Sci. USA.* **89**, 5670–5674 (1992).
- Nicolelis, M. A. L. et al. Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science* **268**, 1353–1358 (1995).
- Donoghue, J. P. et al. Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements. *J. Neurophysiol.* **79**, 159–173 (1998).
- Seidemann, E. et al. Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J. Neurosci.* **16**, 752–768 (1996).
- Riehle, A. et al. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* **278**, 1950–1953 (1997).
- Hatsopoulos, N. G. et al. Information about movement direction obtained from synchronous activity of motor cortical neurons. *Proc. Natl Acad. Sci. USA.* **95**, 15706–15711 (1998).
- Maynard, E. M. et al. Neuronal interactions improve cortical population coding of movement direction. *J. Neurosci.* **19**, 8083–8093 (1999).
- Almaric, M. & Koob, G. F. Depletion of dopamine in the caudate nucleus but not in the nucleus accumbens impairs reaction time performance in rats. *J. Neurosci.* **7**, 2129–2134 (1987).
- Kohonen, T. *Self-organizing Maps* (Springer, New York, 1997).
- Nicolelis, M. A. L., Stambaugh, C. R., Brisben, A. & Laubach, M. in *Methods for Simultaneous Multisite Neural Ensemble Recordings in Behaving Primates* (ed. Nicolelis, M. A. L.) 121–156 (CRC, Boca Raton, 1999).

19. Krippendorff, K. *Information Theory: Structural Models for Qualitative Data* (Sage, Thousand Oaks, California, 1986).
20. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap*. (Chapman & Hall, New York, 1994).
21. Lee, T. W., Girolami, M. & Sejnowski, T. J. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput.* **11**, 417–441 (1999).
22. Laubach, M., Shuler, M. & Nicolelis, M. A. L. Independent component analyses for quantifying neuronal ensemble interactions. *J. Neurosci. Meth.* **94**, 141–154 (1999).
23. Nicolelis, M. A. L. *et al.* Reconstructing the engram: simultaneous, multiple site, many single neuron recordings. *Neuron* **18**, 529–537 (1997).
24. Loeb, G. E. & Gans, C. *Electromyography for Experimentalists* (Univ. of Chicago Press, Chicago, 1986).
25. Nicolelis, M. A. L. *et al.* Simultaneous encoding of tactile information by three primate cortical areas. *Nature Neurosci.* **1**, 621–630 (1998).
26. Ghazanfar, A. A., Stambaugh, C. R. & Nicolelis, M. A. L. Multiple strategies for encoding tactile information by somatosensory thalamocortical ensembles. *J. Neurosci.* (in the press).
27. Mosteller, F. & Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics* (Addison-Wesley, Reading, Massachusetts, 1977).
28. Buckheit, J. & Donoho, D. Improved linear discrimination using time-frequency dictionaries. *Proc. SPIE* **2569**, 540–551 (1995).
29. Reymont, R. A. & Jöreskog, K. G. *Applied Factor Analysis in the Natural Sciences* (Cambridge Univ. Press, New York, 1993).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. Beck, D. Cohen, D. Katz, D. Krupa, M. Shuler and B. Storey-Laubach for comments on the manuscript. The work was supported by grants from the National Institute of Health (M.L. and M.A.L.N.), the National Science Foundation, the Defense Advanced Research Project Agency, the Office of Naval Research, and the Human Frontiers and Whitehall Foundations (M.A.L.N.). J.W. was supported by the Swedish Foundation for International Cooperation in Research and Higher Education and by the Swedish Medical Research Council.

Correspondence and requests for materials should be addressed to: M.L. (e-mail: laubach@neuro.duke.edu).

An electroneutral sodium/bicarbonate cotransporter NBCn1 and associated sodium channel

Inyeong Choi*, Christian Aalkjaer†, Emile L. Boulpaep* & Walter F. Boron*

* Department of Cellular and Molecular Physiology, Yale University School of Medicine, New Haven, Connecticut 06520, USA

† Department of Physiology and Danish Biomembrane Research Centre, University of Aarhus, Aarhus, DK-8000, Denmark

Two electroneutral, Na^+ -driven HCO_3^- transporters, the Na^+ -driven $\text{Cl}^-/\text{HCO}_3^-$ exchanger and the electroneutral $\text{Na}^+/\text{HCO}_3^-$ cotransporter, have crucial roles in regulating intracellular pH in a variety of cells, including cardiac myocytes^{1,2}, vascular smooth-muscle^{3,4}, neurons⁵ and fibroblasts⁶; however, it is difficult to distinguish their Cl^- dependence in mammalian cells. Here we report the cloning of three variants of an electroneutral $\text{Na}^+/\text{HCO}_3^-$ cotransporter, NBCn1, from rat smooth muscle. They are 89–92% identical to a human skeletal muscle clone⁷, 55–57% identical to the electrogenic NBCs and 33–43% identical to the anion exchangers⁸. When expressed in *Xenopus* oocytes, NBCn1-B (which encodes 1,218 amino acids) is electroneutral, Na^+ -dependent and HCO_3^- -dependent, but not Cl^- -dependent. Oocytes injected with low levels of NBCn1-B complementary RNA exhibit a Na^+ conductance that 4,4-diisothiocyanatostilbene-2,2'-disulphonate stimulates slowly and irreversibly.

We designed degenerate primers to identify NBC-related complementary DNAs from rat aorta based on an amino-acid sequence

comparison of several NBCs^{9–12}, a *Caenorhabditis elegans* expressed sequence tag (EST) clone (GenBank accession number Z75541) and rat anion exchangers¹³. We cloned a 294-base pair (bp) polymerase chain reaction (PCR) product, the deduced amino-acid sequence of which is 76% identical to electrogenic renal and pancreatic NBCs. We obtained the 5' end of the open reading frame by 5' rapid amplification of cDNA ends (RACE) of aorta RNA, and obtained the 3' end by performing PCR on a pulmonary-artery cDNA library.

To test whether our three cDNA fragments represent a single transcript, we performed PCR with reverse transcription (RT-PCR) using total RNA from rat aorta and primers designed to amplify the entire coding region. We obtained three distinct, full-length clones (NBCn1-B, C and D). Compared with the NBCn1-D (Fig. 1a), NBCn1-B lacks a 36-residue 'B' domain near the carboxy terminus, and NBCn1-C lacks a 14-residue 'A' domain near the amino terminus (Fig. 1b). The rat NBCn1s are 89–92% identical to a human skeletal-muscle cDNA ('NBC-3/Pushkin')⁷, which we propose to call NBCn1-A. Figure 1a compares the deduced amino-acid sequence of NBCn1-D with the electrogenic rat-kidney NBC (57% identity)¹⁰ and rat AE2 (33–43% identity)¹³. Figure 1c shows the relationships among selected members of the HCO_3^- transporter superfamily. The identity between the NBCn1s and electrogenic NBCs is a fairly uniform 62% along the membrane-spanning regions, but falls off sharply (< 20%) at the amino and carboxy termini. The hydropathy plot of NBCn1-D is similar to that of a typical electrogenic NBC (Fig. 1d).

A northern blot of rat tissues (Fig. 2a), using a probe corresponding to a region common to the NBCn1s, shows substantial hybridization to a 7.5-kilobase (kb) transcript in spleen and a 4.1-kb transcript in testis. The signals are moderate in several other tissues, but undetectable in skeletal muscle. Using primers based on conserved NBCn1 sequences, we obtained a PCR product of the predicted size from spleen, aorta and brain (Fig. 2b).

To test the functional properties, we injected NBCn1-B cRNA into oocytes, and used microelectrodes to monitor membrane potential (V_m) and intracellular pH (pH_i). Oocytes expressing NBCn1-B differ in two principal ways from controls. First, the resting V_m was -30 ± 1 mV ($n = 32$), ~ 30 mV more positive ($P < 0.001$) than in control oocytes (-60 ± 3 mV, $n = 9$). Second, the average initial pH_i was 7.46 ± 0.01 ($n = 29$), 0.12 units higher ($P < 0.01$) than in controls (7.34 ± 0.03 , $n = 5$). Thus, NBCn1-B mediates net HCO_3^- influx with the trace amounts of HCO_3^- present in the incubation medium.

In control oocytes injected with water (Fig. 3a), applying 1.5% $\text{CO}_2/10$ mM HCO_3^- caused a rapid pH_i fall, followed by a long period of stability. In oocytes injected with NBCn1-B cRNA (Fig. 3b), adding $\text{CO}_2/\text{HCO}_3^-$ to the bath caused the usual pH_i fall, followed by a rise owing to HCO_3^- uptake. More importantly, $\text{CO}_2/\text{HCO}_3^-$ did not affect V_m . In oocytes expressing an electrogenic NBC, $\text{CO}_2/\text{HCO}_3^-$ elicits a rapid hyperpolarization, which may be as large as 80 mV (refs 9,10,12). Thus, NBCn1-B mediates electroneutral $\text{Na}^+/\text{HCO}_3^-$ cotransport (that is, $\text{Na}^+:\text{HCO}_3^-$ stoichiometry of 1:1). Removing Na^+ from the bath converted the pH_i recovery to an acidification (Fig. 3b), consistent with reversal of electroneutral $\text{Na}^+/\text{HCO}_3^-$ cotransport. Note that Na^+ removal hyperpolarized oocytes expressing NBCn1-B by 25 ± 1 mV ($n = 10$), much larger than the 11 ± 1 mV ($n = 9$) in controls ($P < 0.001$).

When we acidified oocytes expressing NBCn1-B with butyric acid rather than CO_2 (Fig. 3c), pH_i failed to recover in the presence of Na^+ , and failed to acidify in the absence of Na^+ . Thus, transport by NBCn1-B requires HCO_3^- in both the inward and outward directions. Note also that Na^+ removal in butyrate caused the same large hyperpolarization as in $\text{CO}_2/\text{HCO}_3^-$.

To test whether NBCn1-B encodes a Na^+ -driven $\text{Cl}^-/\text{HCO}_3^-$ exchanger, we removed bath Cl^- in the presence of $\text{CO}_2/\text{HCO}_3^-$ (data not shown); however, Cl^- removal did not speed the pH_i

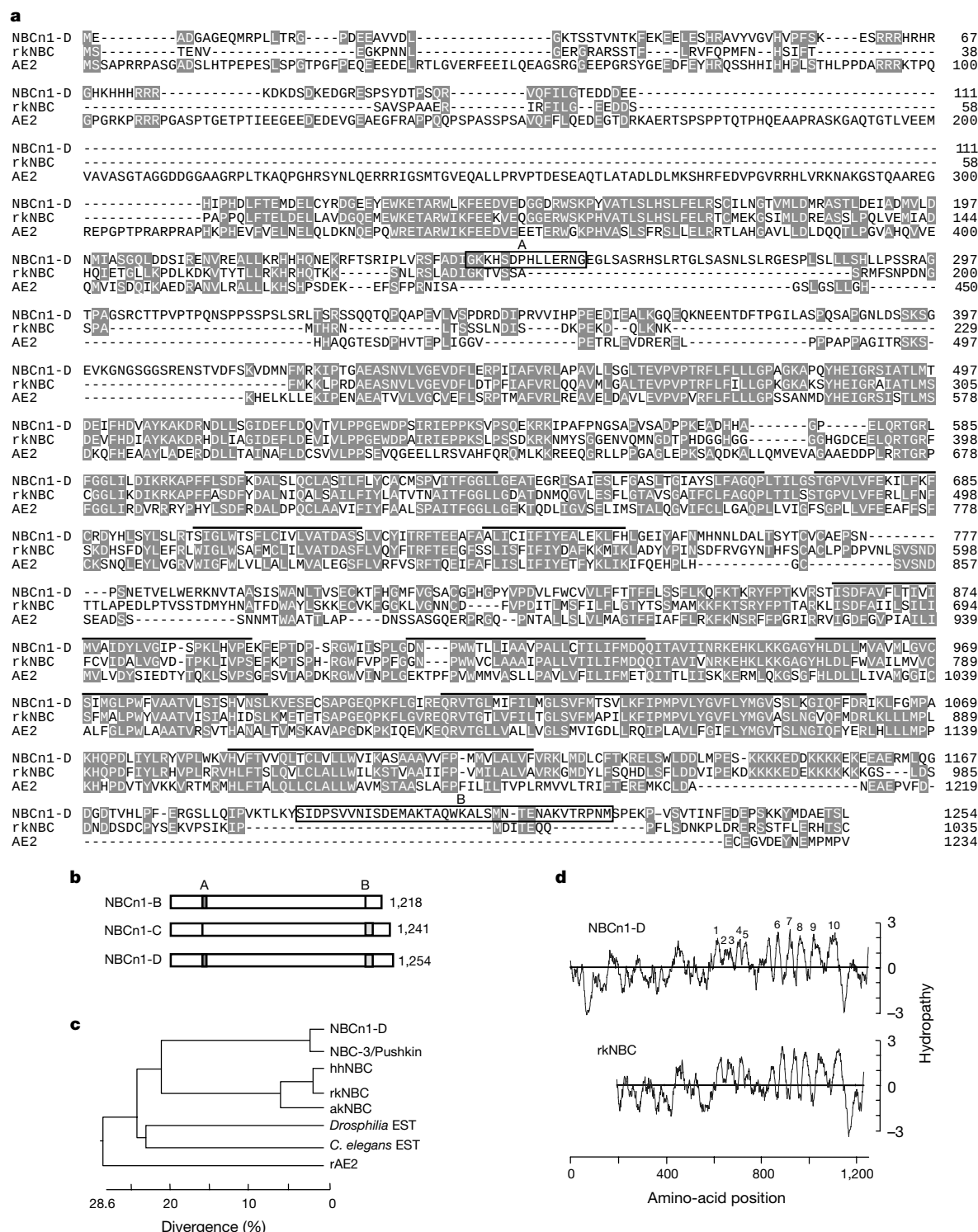


Figure 1 Sequence analysis of NBCn1. **a**, Comparison of amino-acid sequences of NBCn1-D, an electrogenic $\text{Na}^+/\text{HCO}_3^-$ cotransporter from rat kidney (rkNBC) and a rat $\text{Cl}^-/\text{HCO}_3^-$ exchanger (AE2). Residues identical to NBCn1-D are in reverse type. Boxes labelled A and B indicate amino-acid cassettes present or absent in NBCn1-B, NBCn1-C and NBCn1-D. NBCn1-D has three consensus sites for *N*-glycosylation (N780, N790, N800), 14 for protein kinase C, and two for protein kinase A (T230, S246). NBCn1 has a DIDS covalent-reaction motif²⁶ (KLKK at 948–951), which corresponds to the second of two such motifs in electrogenic NBCs; however, the first motif, which is present in AEs and electrogenic NBCs, is disrupted in NBCn1 (KLFLH at 746–749). NBCn1 each has a

putative PDZ-binding motif (ETSL at the C terminus)^{27,28}. **b**, Three variants of NBCn1. **c**, Dendrogram of selected members of HCO_3^- transporter superfamily. Divergence (at the amino-acid level) is indicated by the total length of the horizontal line segments. The GenBank accession numbers for the NBC-related EST clones are Z75541 (*C. elegans*), and AA567741 (*Drosophila*). **d**, Comparison of the hydropathy plots of NBCn1-D (upper panel) and rkNBC (lower panel), using a Kyte–Doolittle algorithm (17 window size). On the basis of the similar hydropathy plots of an electroneutral NBCn1-D and an electrogenic rkNBC, we propose that NBCn1 has at least 10 membrane-spanning segments.

recovery, as might have been expected if NBCn1-B mediated a Cl^- efflux when moving HCO_3^- inward. We then tested whether removing bath Cl^- inhibits NBCn1-B when it moves HCO_3^- outward. Figure 3b shows that NBCn1-B reverses (that is, pH_i decreases) when we remove Na^+ in the presence of $\text{CO}_2/\text{HCO}_3^-$. Removing bath Cl^- did not, however, slow this acidification (Fig. 3d). Using a Na^+ -sensitive microelectrode, we found that oocytes expressing NBCn1-B had an average initial intracellular concentration of Na^+ ($[\text{Na}^+]_i$) of $39.7 \pm 7.3 \text{ mM}$ ($n = 7$), far higher ($P < 0.01$) than that of control oocytes ($6.5 \pm 1.1 \text{ mM}$, $n = 9$). With NBCn1-B operating in the outward direction in the absence of Na^+ , we found that $[\text{Na}^+]_i$ continued to fall during Cl^- removal (Fig. 3e). Moreover, $[\text{Cl}^-]_i$ did not change during Na^+ removal (data not shown). Thus, NBCn1-B does not couple Cl^- to Na^+ or HCO_3^- . Removing Ca^{2+} rather than Cl^- had no effect on the pH_i decline during Na^+ removal (data not shown), ruling out reversed $\text{Na}^+/\text{Ca}^{2+}$ exchange. The effects of disulphonate (4,4-diisothiocyanatostilbene-2,2'-disulphonate, DIDS; $500 \mu\text{M}$) were small and variable. Figure 3f shows that DIDS inhibited NBCn1-B only modestly, perhaps reflecting the absence in NBCn1-B of one of the two DIDS-covalent-reaction motifs found in the electrogenic NBCs. The amiloride derivative EIPA ($200 \mu\text{M}$) did not affect the pH_i recovery (Fig. 3f).

Although the cotransport activity of NBCn1-B is electroneutral, three pieces of evidence suggest that NBCn1-B carries current: (1) resting oocytes are depolarized; (2) Na^+ removal causes an exaggerated hyperpolarization (Fig. 3b); and (3) this hyperpolarization does not require cotransport (Fig. 3c). To characterize this new current, we used a two-electrode voltage clamp. Figure 4a shows current records for H_2O -injected and NBCn1-B oocytes. Figure 4b shows the average I - V relationships. The chord conductance, measured between -80 and $+20 \text{ mV}$, was $7.7 \pm 0.1 \mu\text{S}$ in NBCn1-B oocytes, but only $0.56 \pm 0.01 \mu\text{S}$ in controls. Exposing oocytes to an out-of-equilibrium solution¹⁴ containing 33 mM HCO_3^- at $\text{pH} 7.5$ but negligible CO_2 , does not affect the I - V relationship (C. Severson and I. Grichtchenko, personal communication), indicating that the current neither depends on nor carries HCO_3^- . Removing bath Na^+ decreased inward current by $\sim 60\%$ at -160 mV , and also negatively shifted reversal potential (E_{rev}) by $22.8 \pm 0.6 \text{ mV}$ (Fig. 4c); however, removing other ions had no significant effect on E_{rev} . Figure 4d shows that E_{rev} shifts by $32.4 \pm 4.0 \text{ mV}$ per decade change in the intracellular concentration of Na^+ ($[\text{Na}^+]_i$) indicating that Na^+ carries $\sim 56\%$ of the new current in oocytes expressing NBCn1-B. Ion-selectivity experiments show that the permeability sequence is $\text{Na} > \text{K} = \text{Cs} = \text{N-methyl-D-glucamine}$

(NMDG) (data not shown). We could not compute the relative permeability ratios based on Goldman-Hodgkin-Katz equation because $\sim 44\%$ of the NBCn1-dependent current remained even in the absence of all ions except 1 mM MgCl_2 and HEPES buffer. The remaining inward current probably represents the efflux of an anion.

Surprisingly, DIDS ($500 \mu\text{M}$) increased the inward NBCn1-B-dependent current, with little effect on E_{rev} (Fig. 4e). Na^+ removal caused the inward current to fall to the same low level in the absence

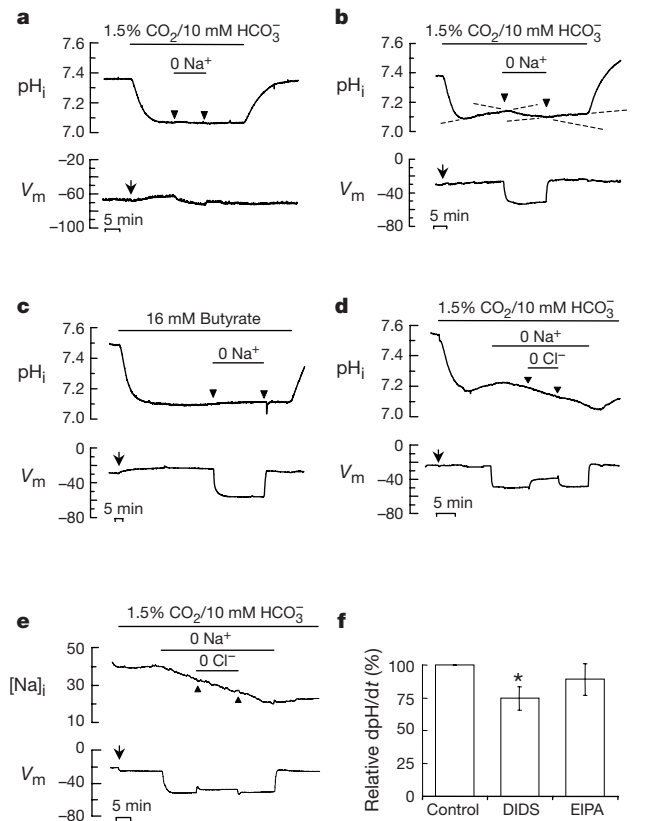


Figure 3 Expression of NBCn1-B in *Xenopus* oocytes. **a**, Water-injected control oocytes. The mean ΔpH_i caused by switching the solution from ND96 to $1.5\% \text{ CO}_2/10 \text{ mM HCO}_3^-$ ($\text{pH} 7.5$) was -0.25 ± 0.02 ($n = 5$). Arrowheads represent the application and removal of the 0-Na^+ solution; during the flat portion of the record, $\text{dpH}/\text{dt} = 0.8 (\pm 1.2) \times 10^{-5} \text{ pH s}^{-1}$. Arrow indicates the time of $\text{CO}_2/\text{HCO}_3^-$ application. V_m is indicated as millivolt (mV). **b**, NBCn1-B-expressing oocytes. The protocol was the same as in **a**. The mean ΔpH_i caused by CO_2 was -0.27 ± 0.01 ($n = 24$). During pH_i recovery in the presence of Na^+ , $\text{dpH}/\text{dt} = 9.4 (\pm 0.9) \times 10^{-5} \text{ pH s}^{-1}$, which is significantly faster than in **a** ($P < 0.001$). During Na^+ removal, $\text{dpH}/\text{dt} = -15.5 (\pm 6.5) \times 10^{-5} \text{ pH s}^{-1}$, which is significantly different from the preceding rate of pH_i increase ($P < 0.01$). **c**, Acidification with butyric acid in oocytes expressing NBCn1-B. The protocol was the same as in **b**, except that the oocyte was acidified by a nominally $\text{CO}_2/\text{HCO}_3^-$ -free solution ($\text{pH} 7.5$) containing 16 mM total butyrate. Arrow indicates the time of butyrate application. During the flat portion of the record, $\text{dpH}/\text{dt} = -2 (\pm 1.26) \times 10^{-5} \text{ pH s}^{-1}$ ($n = 5$), which is significantly different from both the pH_i increase ($P < 0.001$) and decrease ($P < 0.05$) in **b**. **d**, Effect of Cl^- removal on pH_i decay in 0-Na^+ . In the presence of $1.5\% \text{ CO}_2/10 \text{ mM HCO}_3^-$, we first removed Na^+ and then Na^+ and Cl^- together ($n = 3$). **e**, Measurement of $[\text{Na}^+]_i$. The protocol was the same as in **d**, except that we measured $[\text{Na}^+]_i$ instead of pH_i . The $d[\text{Na}^+]_i/\text{dt}$ during Na^+ removal was $-7.1 \pm 1.8 \mu\text{M s}^{-1}$ in NBCn1-B oocytes ($n = 4$), and $-2.2 \pm 0.3 \mu\text{M s}^{-1}$ in H_2O -injected controls ($n = 5$). **f**, Sensitivity to drugs. The dpH/dt was measured during pH_i recovery before and after applying DIDS ($500 \mu\text{M}$) or EIPA ($200 \mu\text{M}$). The relative dpH/dt is the ratio of dpH/dt values in the presence and the absence of the drug ($n = 7$ for DIDS, $n = 4$ for EIPA). The reduction in dpH/dt by DIDS was $25 \pm 9\%$ ($P < 0.05$), whereas the reduction by EIPA was statistically insignificant ($P > 0.1$).

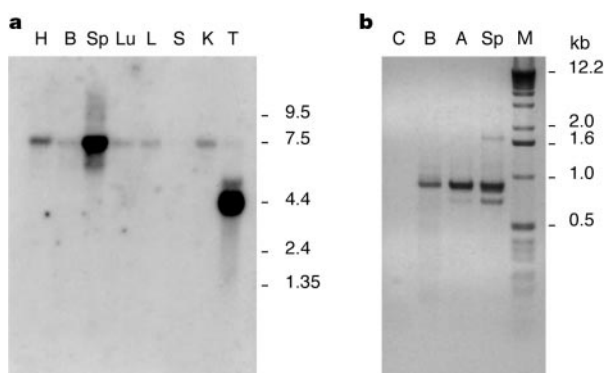


Figure 2 Tissue distribution of NBCn1. **a**, Northern blot analysis of rat tissues probed with a cDNA fragment corresponding to nucleotides 2,068–2,904 of the coding region of NBCn1-B. H, heart; B, brain; Sp, spleen; Lu, lung; Li, liver; S, skeletal muscle; K, kidney; T, testis. **b**, RT-PCR analysis. Primers were designed to amplify the 908-bp fragment conserved among NBCn1s (see Methods). C, control water; B, Brain; A, aorta; Sp, spleen; M, molecular weight markers.

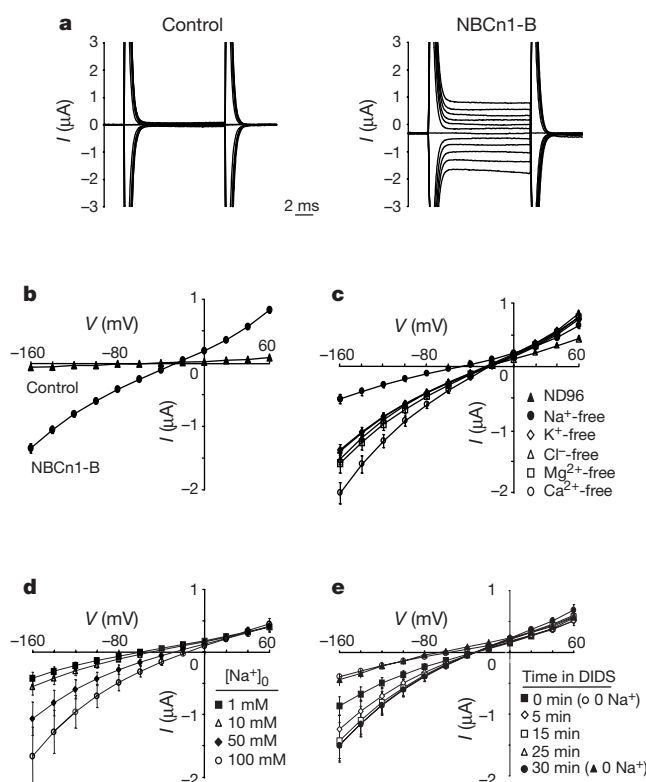


Figure 4 Oocyte currents in the nominal absence of HCO_3^- . **a**, Representative current records in a control (left) and an oocyte expressing NBCn1-B (right). Oocytes were clamped initially at the holding potential of -60 mV, and then stepped from -160 to $+60$ mV, in steps of 20 mV, for 20 ms each. **b**, I - V plots for control and oocytes expressing NBCn1-B. The plots were derived from steady-state currents, such as those in **a** ($n = 26$ for oocytes expressing NBCn1-B, $n = 7$ for water-injected control). **c**, Effect of single-ion substitutions on the I - V relationship. One to two minutes after switching from ND96 to an identical solution lacking one ion, we subjected the oocyte to a voltage-step protocol. NMDG replaced Na^+ . **d**, Effect of altering $[\text{Na}^+]_o$ on the I - V relationship. The

solutions contained 1 , 10 , 50 or 100 mM Na^+ , plus 1 mM MgCl_2 , 5 mM HEPES and sufficient mannitol to maintain osmolality at 210 mOsm. We measured the I - V curves 1 – 2 min after switching to a new solution ($n = 5$). In ion-selectivity experiments (not shown), we switched from 0 to 10 mM NMDG, Na^+ , K^+ or Cs^+ . **e**, Effect of DIDS on the I - V relationship ($n = 6$). We measured I - V relationships before, and 5 – 30 min after applying 500 μM DIDS. Even after more than 30 -min of DIDS treatment, Na^+ removal (filled triangle) still lowered current to values indistinguishable from those in the absence of DIDS (open circle). In control oocytes, DIDS had no effect on the current.

and presence of DIDS (Fig. 4e). DIDS had no effect on water-injected controls ($n = 3$; data not shown). Thus, the DIDS-activated current requires both Na^+ and NBCn1-B. Because DIDS stimulated the current gradually and irreversibly over 15 min, DIDS may interact covalently¹⁵ with NBCn1-B.

The only other examples of transporters with intrinsic channel properties are the glutamate transporters EAAT1 and EAAT3, which have an amino-acid-dependent Cl^- current¹⁶. The Na^+ current of NBCn1 does not resemble currents described previously in *Xenopus* oocytes. It is not the non-selective ($P_{\text{Na}}/P_{\text{K}} \approx 1$) cation current that appears in oocytes injected with high levels of cRNAs (for example, 50 ng oocyte) encoding membrane proteins¹⁷; that current activates very slowly and is inhibited by DIDS. Nor is it the non-selective ($P_{\text{Na}}/P_{\text{K}} \approx 0.67$) cation current activated by maitotoxin¹⁸. Thus, we suggest that the DIDS-activatable Na^+ current flows through NBCn1-B.

Our experiments prove the existence of the long-hypothesized electroneutral $\text{Na}^+/\text{HCO}_3^-$ cotransporter, which is crucial in regulating pH_i in myocardial cells¹, as well as in smooth-muscle cells from rat mesenteric artery³ and guinea-pig ureter⁴. Moreover, NBCn1 may contribute to the background Na^+ conductance in these tissues. Because pH_i changes modulate muscle tension in cardiac muscle¹⁹ and vascular smooth muscle^{20,21}, alterations in NBCn1 activity might be involved in heart failure and myocardial infarction²². Cloning the cDNAs that encode the electroneutral $\text{Na}^+/\text{HCO}_3^-$

cotransporters may lead to new insights into these disease processes, as well as into factors determining the $\text{Na}^+:\text{HCO}_3^-$ stoichiometry of NBCs. Finally, if the Na^+ conductance is part of the cotransport pathway, studying this conductance may for the first time provide insight into the nature of the pathway. □

Methods

Cloning of NBCn1

We cloned NBCn1 in three sections. The central piece was a PCR product obtained from tRNA isolated from rat aorta tissue²³. Degenerate primers were designed to amplify fragments corresponding to the putative transmembrane domains 8 and 9 of NBCs. The upstream primer was $5'$ -AARGGKICGGITWYCACTIGAY- $3'$ (nucleotide positions 2,447–2,470 of rkNBC); the downstream primer was $5'$ -RTGIGGITGKTKYTTIAKIGGCAT- $3'$ (2,795–2,818 of rkNBC). In the first three cycles of PCR, we performed denaturation at 94°C for 30 s, annealing at 42°C for 1 min and extension at 72°C for 2 min. In the next 20 cycles, the annealing temperature was 50°C . We then diluted the product 200 -fold with water, and performed the second-round PCR with the same upstream primer used for the first round of PCR and a nested downstream primer ($5'$ -IGAIGIAYICCATRTADAGRAA- $3'$, 2,729–2,752 of rkNBC).

The $5'$ section was obtained by $5'$ RACE of RNA from rat aorta. The $3'$ section was obtained by performing PCR on a rat pulmonary artery cDNA library (Stratagene) using a gene-specific sense primer and a T7 vector primer adjacent to the cDNA insert sites. We established the final sequence by comparing 29 overlapping PCR products (as well as a 1.5 -kb clone from the rat pulmonary artery cDNA), and obtaining a consensus at each nucleotide position. The full-length cDNA used in our experiments was obtained by ligating error-free PCR products. We also performed PCR on rat-aorta RNA and obtained full-length cDNA clones.

Tissue distribution

A rat multiple-tissue northern blot (catalogue number 7764-1) was purchased from Clontech (Palo Alto, CA). The cDNA fragment corresponding to nucleotide 2,068–2,904 of the coding region of NBCn1-B was random-primed. Hybridization was done at 68 °C for 2 h in the ExpressHyb hybridization buffer (Clontech), at a probe concentration of 3×10^6 c.p.m. ml⁻¹. The filter was washed according to the company's manual. For RT-PCR, the upstream primer was 5'-GCATTTGTGAGACTGGCTCCTGCAGTTCTC-3' (1,306–1,335 of NBCn1-B); the reverse primer was 5'-TCCTAAGTGAAAGAGTTTCTCCAAG GCTTC-3' (2,185–2,214 of NBCn1-B). These primers are expected to amplify a 908-bp PCR product, which is conserved among NBCn1s. Total RNAs were isolated from rat spleen, brain and aorta. After reverse transcription, PCR was performed for 35 cycles with denaturation at 94 °C for 45 s, annealing at 60 °C for 1 min, and extension at 68 °C for 2 min with a 20-s extended increment.

Oocytes

The NBCn1-B cDNA (–22 to +3,852 of the coding region) was inserted into the oocyte expression vector pGH19 (ref. 24). *In vitro* transcription as well as injection of cRNA (10–17 ng per oocyte) into *Xenopus laevis* oocytes were prepared as described¹². Oocytes were maintained in OR3 buffer⁹ for 3–7 d at 18 °C before use. The CO₂/HCO₃⁻-free ND96 contained (in mM) 96 NaCl, 2 KCl, 1 MgCl₂, 1.8 CaCl₂ and 5 HEPES (pH 7.5). In 1.5% CO₂/HCO₃⁻-equilibrated solutions, 10 mM NaHCO₃ replaced 10 mM NaCl. In 0-Na⁺ solutions, choline or NMDG replaced Na⁺. In 0-Cl⁻ solutions, gluconate replaced Cl⁻. Na⁺ was used to replace K⁺, Ca²⁺ and Mg²⁺.

V_m and ion-selective microelectrodes

The microelectrodes were prepared as described⁹. The pH electrode tip was filled with proton ionophore 1 cocktail B (Fluka), and back-filled with a phosphate buffer (pH 7.0)²⁵. The Na⁺ or Cl⁻ electrodes were filled with sodium ionophore cocktail A (Fluka) or chloride ionophore cocktail (Fluka), and calibrated in solutions containing NaCl or KCl (5, 10, 20, 40 and 80 mM), respectively. The electrodes were connected to a high-impedance electrometer (model FD-223; WPI, Inc., Sarasota, FL), which was in turn connected to the A-D converter of a computer.

Voltage-clamping

We used a two-electrode voltage clamp (model OC 725A, Warner Instruments, New Haven, CT), connected to a Digidata2000 interface (Axon Instruments, Foster City, CA). Data were acquired and analysed using pClamp 6 software (Axon Instruments). All voltage-clamp experiments were performed in the nominal absence of HCO₃⁻. Statistical significance was judged from Student's *t*-tests. Data are expressed as mean $\pm$ s.e.m.

Received 15 December 1999; accepted 4 April 2000.

- Dart, C. & Vaughan-Jones, R. D. Na⁺-HCO₃⁻ symport in the sheep cardiac Purkinje fibre. *J. Physiol.* **451**, 365–385 (1992).
- Lagadic-Gossmann, D., Buckler, K. J. & Vaughan-Jones, R. D. Role of bicarbonate in pH recovery from intracellular acidosis in the guinea-pig ventricular myocyte. *J. Physiol.* **458**, 361–384 (1992).
- Aalkjaer, C. & Hughes, A. Chloride and bicarbonate transport in rat resistance arteries. *J. Physiol.* **436**, 57–73 (1991).
- Aickin, C. C. Regulation of intracellular pH in the smooth muscle of guinea-pig ureter: Na⁺ dependence. *J. Physiol.* **479**, 301–316 (1994).
- Schwiening, C. J. & Boron, W. F. Regulation of intracellular pH in pyramidal neurons from the rat hippocampus by Na⁺-dependent Cl⁻-HCO₃⁻ exchange. *J. Physiol.* **475**, 59–67 (1994).
- L'Allemand, G., Paris, S. & Pouyssegur, J. Role of a Na⁺-dependent Cl⁻/HCO₃⁻ exchange in regulation of intracellular pH in fibroblasts. *J. Biol. Chem.* **260**, 4877–4883 (1985).
- Pushkin, A. *et al.* Cloning, tissue distribution, genomic organization, and functional characterization of NBC3, a new member of the sodium bicarbonate cotransporter family. *J. Biol. Chem.* **274**, 16569–16575 (1999).
- Kopito, R. R. & Lodish, H. F. Primary structure and transmembrane orientation of the murine anion exchange protein. *Nature* **316**, 234–238 (1985).
- Romero, M. F., Hediger, M. A., Boulpaep, E. L. & Boron, W. F. Expression cloning and characterization of a renal electrogenic Na⁺/HCO₃⁻ cotransporter. *Nature* **387**, 409–413 (1997).
- Romero, M. F., Fong, P., Berger, U. V., Hediger, M. A. & Boron, W. F. Cloning and functional expression of rNBC, an electrogenic Na⁺-HCO₃⁻ cotransporter from rat kidney. *Am. J. Physiol.* **274**, F425–F432 (1998).
- Abuladze, N. *et al.* Molecular cloning, chromosomal localization, tissue distribution, and functional expression of the human pancreatic sodium bicarbonate cotransporter. *J. Biol. Chem.* **273**, 17689–17695 (1998).
- Choi, I., Romero, M. F., Khandoudi, N., Bril, A. & Boron, W. F. Cloning and characterization of a human electrogenic Na⁺-HCO₃⁻ cotransporter isoform (hhNBC). *Am. J. Physiol.* **276**, C576–C584 (1999).
- Kudrycki, K. E., Newman, P. R. & Shull, G. E. cDNA cloning and tissue distribution of mRNAs for two proteins that are related to the Band 3 Cl-HCO₃⁻ exchanger. *J. Biol. Chem.* **265**, 462–471 (1990).
- Zhao, J., Hogan, E. M., Bevensee, M. O. & Boron, W. F. Out-of-equilibrium CO₂/HCO₃⁻ solutions and their use in characterizing a new K/HCO₃ cotransporter. *Nature* **374**, 636–639 (1995).
- Cabantchik, Z. I., Knauf, P. A. & Rothstein, A. The anion transport system of the red blood cell. The role of membrane protein evaluated by the use of 'probes'. *Biochim. Biophys. Acta* **515**, 239–302 (1978).
- Wadiche, J. I., Amara, S. G. & Kavanaugh, M. P. Ion fluxes associated with excitatory amino acid transport. *Neuron* **15**, 721–728 (1995).
- Tzounopoulos, T., Maylie, J. & Adelman, J. P. Induction of endogenous channels by high levels of heterologous membrane proteins in *Xenopus* oocytes. *Biophys. J.* **69**, 904–908 (1995).
- Biefeld-Ackermann, A., Range, C. & Korbacher, C. Maitotoxin (MTX) activates a nonselective cation channel in *Xenopus laevis* oocytes. *Pflügers Arch.* **436**, 329–337 (1998).

- Orchard, C. H. & Kentish, J. C. Effects of changes of pH on the contractile function of cardiac muscle. *Am. J. Physiol.* **258**, C967–C981 (1990).
- Aalkjaer, C. & Poston, L. Effects of pH on vascular tension: which are the important mechanisms? *J. Vasc. Res.* **33**, 347–359 (1996).
- Smith, G. L., Austin, C., Crichton, C. & Wray, S. A review of the actions and control of intracellular pH in vascular smooth muscle. *Cardiovasc. Res.* **38**, 316–331 (1998).
- Yu, M., Sandmann, S. & Unger, T. Angiotensin AT₂ receptor mediates expression of Na⁺-HCO₃⁻ symporter in rat myocardium after experimental myocardial infarction. *Hypertension* **32**, 807 (1998).
- Sambrook, J., Fritsch, E. F. Maniatis, T. *Molecular Cloning: a Laboratory Manual* 2nd edn (Cold Spring Harbor Press, Cold Spring Harbor, 1989).
- Trudeau, M. C., Warmke, J. W., Ganetzky, B. & Robertson, G. A. HERG, a human inward rectifier on the voltage-gated potassium channel family. *Science* **269**, 92–95 (1995).
- Siebens, A. W. & Boron, W. F. Effect of electroneutral luminal and basolateral lactate transport on intracellular pH in salamander proximal tubules. *J. Gen. Physiol.* **90**, 799–831 (1987).
- Kopito, R. R. *et al.* Regulation of intracellular pH by a neuronal homolog of the erythrocyte anion exchanger. *Cell* **59**, 927–937 (1989).
- Kim, E., Niethammer, M., Rothschild, A., Jan, Y. N. & Sheng, M. Clustering of Shaker-type K⁺ channels by interaction with a family of membrane-associated guanylate kinases. *Nature* **378**, 85–88 (1995).
- Short, D. B. *et al.* An apical PDZ protein anchors the cystic fibrosis transmembrane conductance regulator to the cytoskeleton. *J. Biol. Chem.* **273**, 19797–19801 (1998).

Acknowledgements

We thank B. Schmitt, M. Bevensee and M. Romero for valuable discussions. We also thank D. Wong for computer support. This work was supported by NIH grants (W.F.B.) and by the Danish Council (C.A.). I.C. was supported by a postdoctoral fellowship from the American Heart Association.

Correspondence and requests for materials should be addressed to W.F.B. (e-mail: walter.boron@yale.edu). Sequences have been deposited in GenBank under accession numbers AF070475, AF069511 and AF080106 for NBCn1-B, NBCn1-C and NBCn1-D, respectively.

Cyclin D control of growth rate in plants

Claire E. Cockcroft*, Bart G. W. den Boer*†, J. M. Sandra Healy*‡ & James A. H. Murray*

* Institute of Biotechnology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QT, UK

† Aventis CropScience N.V., Jozef Plateastraet 22, B-9000, Gent, Belgium

The mechanisms by which plants modulate their growth rate in response to environmental and developmental conditions are unknown, but are presumed to involve specialized regions called meristems where cell division is concentrated^{1–5}. The possible role of cell division in influencing meristem activity and overall plant growth rate is controversial, with a prevailing view that cell division is secondary to higher order meristem controls^{1,2,6,7}. Here we show that a reduction in the length of the cell-cycle G1 phase and faster cell cycling occur when the rate of cell division in transgenic tobacco plants is increased by the plant D-type cyclin CycD2 (ref. 8). The plants have normal cell and meristem sizes, but elevated overall growth rates, an increased rate of leaf initiation and accelerated development in all stages from seedling to maturity. We conclude that cell division is a principal determinant of meristem activity and overall growth rate, and propose that modulation of plant growth rate is achieved through regulation of G1.

The G1 phase of the cell cycle is the primary control point in yeast, animals and plants^{9–11}, and plant D-type cyclins (CycD) are suggested to control both commitment to cell division and responses of plant cells to extracellular signals during G1^{8,11,12}. Cyclins of the CycD3 group have a role in S-phase entry in response to plant hormones and spatial signals¹², whereas CycD2 is activated earlier in

‡ Present address: Dipartimento di Genetica, IV Piano, Torre A, Via Celoria, 26, Università di Milano, 20133 Milano, Italy

G1 and responds to sugar availability³⁰. The *Arabidopsis* gene encoding *CycD2* (*AtCycD2;1*, here *CycD2;At*)⁸ was selected as a candidate upstream regulator of cell division.

Transgenic tobacco plants carrying *CycD2;At* under control of a constitutive cauliflower mosaic virus (CaMV) 35S promoter expressed *CycD2;At* messenger RNA (Fig. 1a) and protein (Fig. 1b). A band detected in protein gel blots of wild-type tobacco extracts by the *CycD2;At* antiserum (Fig. 1b) was not associated with a cyclin-dependent kinase (Cdk) activity, as immunoprecipitation with the anti-*CycD2;At* antiserum, followed by assay of the immunoprecipitates for Cdk activity did not detect any activity in wild-type tobacco extracts. In contrast, strong kinase activity was observed in immunoprecipitates from *CycD2;At*-expressing transgenic plants (Fig. 1c). The tobacco *Cdc2a* protein¹³, a Cdk of the archetypal PSTAIRE class¹⁴, was co-immunoprecipitated only in extracts from tobacco expressing *CycD2;At* (Fig. 1d). No association of *CycD2;At* with the tobacco non-PSTAIRE Cdk *Cdc2b* was observed (data not shown). The *Arabidopsis* *CycD2* protein therefore interacts with tobacco *Cdc2a* to form functional kinase complexes in transgenic tobacco plants.

Vegetative transgenic tobacco plants expressing *CycD2;At* were morphologically similar to wild-type controls and non-transgenic segregant siblings; however, *CycD2* transgenic plants had an accelerated growth rate, observed as an increased rate of above-ground biomass accumulation and more rapid attainment of flowering size. Growth enhancement was observed in young seedlings in soil (Fig. 2a) and *in vitro* (data not shown) and continued during vegetative growth in soil (Fig. 2b–d). Root growth was also enhanced in *CycD2* transgenics (data not shown), showing that both root and shoot meristems are responsive to *CycD2*. Transgenic plants, wild-type segregants and untransformed controls initiated flowering at the same height, and reached the same terminal height (Fig. 2d), but because of the increased growth rate of *CycD2* transgenics, flowering started 9–14 d earlier (Table 1). These effects were observed in ten independent T1 lines tested in growth experiments *in vitro* and in a glasshouse, and were also analysed in over 250 individual T2 plants. In segregating populations, including crosses to untransformed wild-type plants, transgene-free segregants behaved as wild-type controls and growth enhancement was

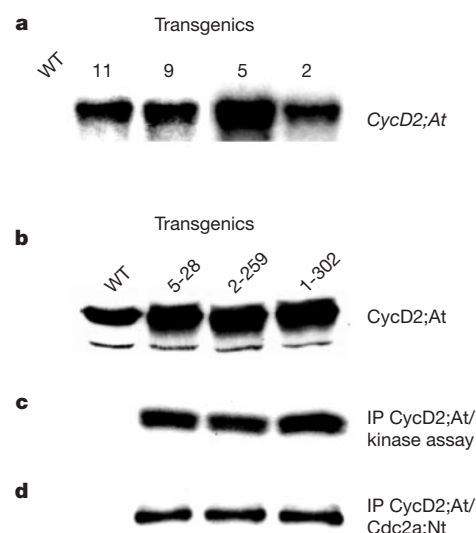


Figure 1 Expression, activity and interaction of *CycD2;At* in tobacco. **a**, RNA gel blot analysis of independent tobacco primary transformants. **b**, Protein gel blot analysis of protein extracts from wild-type (WT) tobacco and individual T1 plants from line 5 and 2 using antiserum raised against *Arabidopsis* *CycD2*. Note cross-reacting protein in wild-type does not have kinase activity or interact with *Cdc2a;Nt* (see **c** and **d**). **c**, Histone H1 kinase activity present in anti-*CycD2;At* immunoprecipitates. **d**, Protein gel blot of anti-*CycD2;At* immunoprecipitates probed with antiserum against a specific C-terminal peptide of *Cdc2a;Nt*.

always correlated with inheritance of the resistance marker carried with the *CycD2;At* transgene (Table 1). Final leaf sizes, internode length and leaf number were unaltered between wild-type and *CycD2;At*-expressing transgenic plants (data not shown).

The faster growth rate observed in *CycD2;At*-expressing transgenics might be a consequence of alterations in meristem activity, or more rapid elongation of leaves once initiated. Measurements of the rate of leaf expansion showed no differences between *CycD2;At*-expressing transgenic plants and wild-type controls (Fig. 3a). Examination of shoot apical meristems by sectioning and scanning electron microscopy revealed no apparent differences in the size or structure of the meristem or of their component cells (Table 2) between transgenic and control wild-type plants, with the exception of peripheral zone cell length which was increased by 7%. This marginal increase in cell size is unlikely to explain the faster growth, as mature leaf cells are not different in size (data not shown). The rate of initiation of new leaves (plastochron), however, was found to be significantly faster in *CycD2;At*-expressing transgenic seedlings and plants (Fig. 3b, c).

Initiation and development of leaf primordia involves both cell proliferation within the primordium and recruitment of cells from the meristem^{1,15}, which must be replenished if the meristem is not to

Table 1 Mean plant height and flowering time in untransformed tobacco and siblings with and without transgene from T1 and T2 populations

Line	Genotype	Height at 37 days (cm)	Height at 55 days (cm)	No. of days to 0.5 cm inflorescence (cm)
Wild-type	No transgene	3.5 ± 0.83	58.2 ± 8.4	72.1 ± 2.3
2 (T1)	hom	7.7 ± 2.2*	72 ± 9.7**	58.5 ± 2.5*
	hem	7.7 ± 2.2*	74.9 ± 9.6*	59.3 ± 3.1*
	No transgene	3.8 ± 1.7	54.5 ± 14.2	73 ± 5.9
5 (T1)	hom + hem	7.1 ± 1.7*	71.2 ± 10.7*	62.7 ± 3.4*
	No transgene	3.9 ± 1.1	53.4 ± 10.7	71.5 ± 5.8
2-3 (T2)	hom	7.7 ± 1.5*	76.3 ± 8.8*	59.4 ± 3.0*
	hem	7.6 ± 1.5*	77.1 ± 9.6*	59.3 ± 1.7*
	No transgene	4.0 ± 1.0	56.8 ± 7.3	73.5 ± 5.4
2-3xwt (T2)	hem	6.9 ± 1.1*	74.7 ± 9.8*	59.4 ± 2.8*
	No transgene	3.7 ± 0.8	61.1 ± 9.6	69.1 ± 2.8
5-304 (T2)	hom	6.6 ± 1.3*	78.3 ± 9.1*	65.8 ± 2.5*
	hem	6.6 ± 1.1*	79.4 ± 6.2*	64.5 ± 3.0*
	No transgene	3.2 ± 1.0	57.8 ± 15.8	74.5 ± 7.8

Untransformed wild-type ($n = 12$) plants were compared with segregating *CycD2;At*-expressing populations ($n = 30$) for each line; values are mean ± s.d. Lines 2 (T1), 5 (T1), 2-3 (T2) and 5-304 (T2) segregate 3:1 for the introduced transgene and the cross 2-3xWT (T2) segregates 1:1. The genotype of individual plants was determined by segregation analysis of selfed seed. T1, seed from primary transformant; T2, seed from T1 plant. First number in designation indicates independent line number and, for T2 seed, second number indicates individual T1 plant from which seed was collected. Significance levels of transgenic classes compared with wild type: *, $P < 0.002$; **, $P < 0.01$; other values are not significant ($P > 0.05$). hom, homozygous, hem, hemizygous.

Table 2 Cell sizes in shoot apical meristems of wild-type and *CycD2;At*-expressing transgenic plants

Line	CZ		PZ	
	Length (μm)	Width (μm)	Length (μm)	Width (μm)
Wild type	18.8 ± 1.8 ($n = 25$)	16.1 ± 3.8 ($n = 25$)	18.0 ± 1.4 ($n = 46$)	10.9 ± 2.5 ($n = 46$)
Transgenic plants (pooled)	19.5 ± 1.4 ($n = 45$) $P = 0.110$	15.7 ± 3.4 ($n = 45$) $P = 0.747$	19.3 ± 2.1 ($n = 102$) $P = 0.002^*$	11.0 ± 2.5 ($n = 102$) $P = 0.870$

Mean ± s.d. length and width are given for central zone (CZ) and peripheral zone (PZ) cells of untransformed (wild-type; 2 plants) and homozygous transgenic lines 2-302 and 5-28 (7 plants). n , number of cells measured; *, significant, $P = 0.002$; t-test.

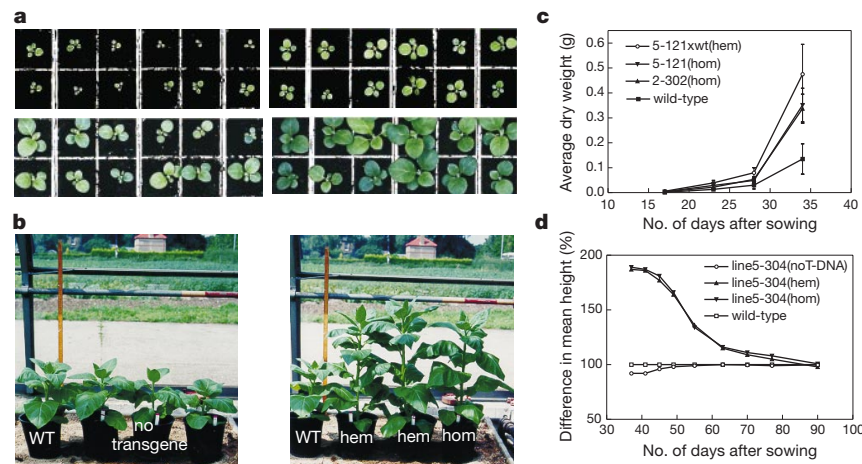


Figure 2 Phenotypic characterization of *CycD2;At*-expressing tobacco plants. **a**, Wild-type (left) and *CycD2;At*-expressing transgenic seedlings (5-121xWT; right) 15 d post-germination (d.p.g.) (top) and 21 d.p.g. (bottom). **b**, Left, wild-type and typical transgene-free segregants of line 3. Right, wild-type and three typical transgene-containing plants from the same population. All plants photographed at 48 d.p.g. hem, hemizygous for

become smaller and eventually depleted of stem cells². As meristem sizes are identical in both control and *CycD2;At*-transgenic plants, the reduced interval between successive primordia implied a faster rate of cell division.

The overall rate of cell division in the meristem depends on the length of the cell cycle and the proportion of cells involved in division. Cells in intact seedlings of *CycD2;At*-expressing transgenic and controls were synchronized using aphidicolin, which blocks cells in early S phase¹⁶. Samples taken after release from the aphidicolin block were monitored for the number of mitotic figures present in the root meristem, and two sequential peaks of mitosis were observed (Fig. 4a). Effects were observed on both the overall length of the cell cycle and the proportion of cells actively involved in division. The interval between the mitotic peaks showed that the cell cycle was reduced from 10.1 h in control wild-type meristems to 9 h in *CycD2;At*-expressing meristems. Because no significant differences were observed between the samples, except for a single time point, until 11 h after release from the block, the length of S, G2 and M phases of the cell cycle after aphidicolin release were unaffected¹⁷. The reduction in the length of this cell cycle was therefore due to a 40% reduction in the G1 phase from 2.7 h in wild-type to 1.6 h in *CycD2;At*-expressing transgenics¹⁸. In addition, 70% more cells entered the second cycle in transgenic meristems (Fig. 4a).

Meristems contain sub-populations of faster and slower dividing cells^{1,19}, and synchronization techniques measure the length of the fastest cell cycles^{1,17}. In *CycD2;At*-expressing transgenic plants, we suggest that the fastest meristem cell-cycle times are reduced because of a shortening of G1 phase. In addition, there is a significantly increased proportion of fast dividing cells, which may be interpreted as a reduction in the length of G1 phase of slow cycling cells normally experiencing an extended G1. G1 is the most variable cell-cycle phase in both root and shoot apical meristem cells in response to developmental^{1,19–21} and environmental^{3–5} signals, and alterations in the size of the fast cycling population have already been proposed as a major mechanism of growth rate control in environmental responses^{3,4}. Analogous to the effects we observed in *CycD2;At*-expressing transgenic tobacco, elevated CO₂ levels increase overall growth in the grass *Dactylis* by shortening G1 phase and increasing the population of fast cycling cells³.

To confirm the faster rate of production of new cells in *CycD2* transgenics, we took unsynchronized meristems and measured the accumulation of metaphase figures after treatment with the mitosis-blocking drug colchicine²². Higher numbers of metaphase figures

were observed at all time points for both root and shoot meristems of *CycD2;At*-expressing seedlings (Fig. 4b), confirming the larger fraction of cells involved in division and indicating that, in agreement with the enhanced growth of both shoots and roots, similar changes in cell division occur in both principal meristems.

Two mechanisms may be proposed to explain the *CycD2* promotion of plant growth. First, *CycD2* could specifically accelerate G1, and the increased rate of cell division would therefore be due to the direct action of *CycD2* on cell-cycle progression. Alternatively, *CycD2* kinase activity might have the function in plants of stimulating cellular growth within the meristem because of an effect on general biosynthesis, which could then increase cell division perhaps as a consequence of reduced time required to reach a critical

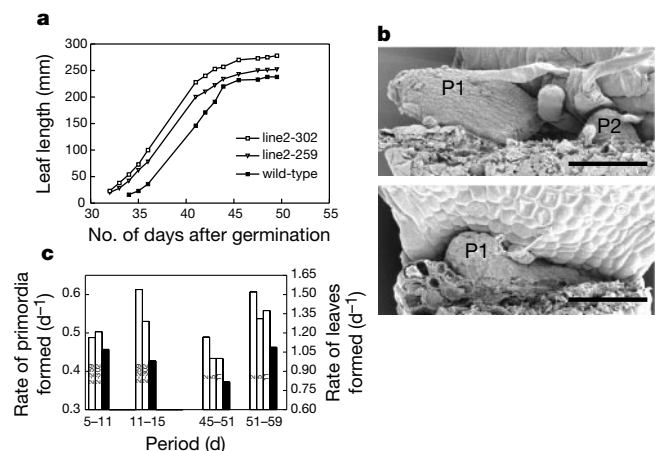


Figure 3 Leaf expansion and initiation in wild-type and *CycD2;At*-expressing transgenic plants. **a**, Expansion rate of leaf 11. Equality-of-slope tests (35–42 d) show no significant differences between wild type and 2-302 ($P = 0.08$), wild type and 2-259 ($P = 0.59$) and 2-302 and 2-259 ($P = 0.23$). **b**, Scanning electron micrograph of a typical wild-type seedling (bottom) and a homozygous transgenic seedling (line 2) (top) at 5 d.p.g. P1, primordium 1. P2, primordium 2. Scale bar, 100 μ m. **c**, Primordium and leaf initiation rates in wild-type control and *CycD2;At*-expressing transgenic seedlings. Black bar, wild type. Rate of primordia formation (left scale) was determined from primordium number of T2 and wild-type seedlings at 5, 11 and 15 d.p.g. using scanning electron microscopy. Rate of leaf formation (right scale) was calculated from inspection of visible leaf primordia at 45, 51 and 59 d.p.g. on T1 plants of lines 2, 5 and 11.

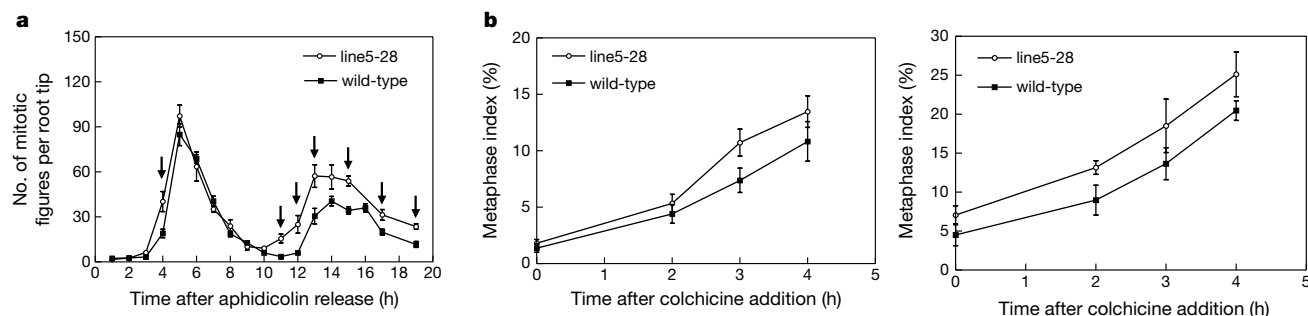


Figure 4 Cell-cycle analysis in wild-type and *CycD2;At*-expressing transgenic meristems. **a**, Transgenic T2 (line 5-28) and wild-type seedlings were synchronized in early S phase by aphidicolin block and release. Progression of root meristem cells through the cell cycle was monitored hourly after release by changes in the number of mitotic figures

(metaphases and anaphases) in the root tip. Arrows denote values significantly different ($P < 0.05$) between transgenic and wild type. **b**, Metaphase index in root tip (left) and in young leaf primordia (right) before and after treating unsynchronized transgenic (line 5-28) and wild-type seedlings with colchicine.

cell size during G1 (ref. 18). In yeasts and animals, increased expression of G1 cyclins has the direct effect of reducing the length of G1, but this normally leads either to a reduction in cell size^{18,23–26} or to compensatory increases in the length of other cell-cycle phases²⁶. Neither effect was observed in *CycD2;At*-expressing tobacco plants, suggesting that new relationships between cell growth and cell division controls may exist in plant meristems.

Experiments have shown that elevated levels of a mitotic (B-type) cyclin can increase root growth, which suggests that G2/M controls may be rate limiting in root meristems²⁷; however, the mechanism by which this occurred is unclear because no link to an increase in Cdk activity or a change in the rate of cell division was shown. Expression of a dominant-negative Cdk has been shown to inhibit cell division leading to increased cell size²⁸, presumably because of the uncoupling of cell growth from division. Our results suggest that *CycD2*-*Cdc2a* kinases act on the G1 phase of the cell cycle to control cell division rate in both shoot and root meristems. Enhanced *CycD2* levels increase the rate of cell division and plant growth, demonstrating that the overall rate of plant growth may be controlled by meristem cell division. Our observations are not consistent with the alternative view that cell division acts in a subordinate role to overall meristem controls in determining growth rate, by serving merely to subdivide cellular space^{1,2,6,7}. Furthermore, we show that faster cell-division rates within the shoot apical meristem do not affect meristem size or structure, but result in the more rapid formation of new primordia. The plastochron thus appears to be determined by the rate of cell division within the meristem, consistent with the requirement to replenish cells lost in the process of initiating a primordium. If cells are more rapidly produced, primordia are able to form at shorter intervals since the critical meristem size for primordium initiation is regained more rapidly¹.

We therefore conclude that overall plant growth rate is directly correlated with cell division, and propose that environmental and developmental controls of growth rate act by regulating *CycD* kinase activity and cell division, thus providing a simple mechanism for the integration of multiple signals affecting the meristem. Finally, we have shown that plant growth rate is not inherently optimized, indicating the potential of *CycD* cyclins as target for crop enhancement. □

Methods

Plant growth analysis

Standard *Agrobacterium*-mediated transformation techniques were used to generate transformants of *Nicotiana tabacum* var. Xanthi carrying the *CycD2* complementary DNA⁸ expressed from the 35S promoter of pART7/27 (ref. 29). Tobacco plants were grown in a glasshouse during summer with a minimum night temperature of 20 °C and supplementary lighting to provide 18 h of light. Young plants were grown at 24 °C with an 18-h photoperiod of fluorescent lighting from Philips TLD50W/840HF tubes providing an intensity of 6,000–10,000 lux at soil level. Seedlings were grown *in vitro* at 23 °C under

24-h fluorescent lighting as above providing 3,500–6,000 lux on Murashige and Skoog (MS) salts (ICN), 1.5% sucrose and 1% bacto-agar (Difco). Dry weight of apical regions was determined after drying 48 h (70 °C). To determine cell size, cells were measured on photographs of thin sections of fixed and embedded apical regions and mature leaves. To determine leaf expansion, the length of leaf 11 was measured with a ruler at the indicated time points. Primordia were scored as $n+1$ when the n th primordium was visible and $n+1$ just discernable. The size of primordium n relative to its maximal size was then used to calibrate intermediate stages. For example, a meristem with a primordium of $x \mu\text{m}$ was scored as $(n + x/y)$, where y is the maximum size of primordium n when primordium $n+1$ appears.

Cell cycle analyses

For root-cell synchronization, tobacco seedlings were germinated *in vitro* in liquid B5 (Gamborgs B5; ICN; with 2% glucose) at 120 r.p.m., 23 °C until 7 days post germination (d.p.g.) and then incubated in 10 μM aphidicolin (24 h), washed four times for 15 min with B5, and placed in fresh B5. Every hour, 6–10 seedlings were fixed, hydrolysed, stained, squashed and metaphases and anaphases were counted (as in ref. 16 but with 2-min hydrolysis). Cell-cycle times were calculated as the interval between the midpoints of the mitotic peaks and the length of the G1 phase, as described¹⁷. For metaphase accumulation, seedlings at 6 d.p.g. as above were incubated in 0.05% colchicine²². At each time point, root and apical parts of 5–10 seedlings were separated and treated as above. Root-tip metaphase index was calculated from the total meristem cell number of 4,385 (average total number of root-tip cells in which metaphases were detected). Apical metaphases were scored in the smallest leaf primordium (~100 μm high) and divided by total cells present to calculate metaphase index.

Protein methods

One gram of plant tissue, ground in liquid nitrogen and resuspended in 1 ml of 50 mM Tris-HCl pH 7.5, 75 mM NaCl, 15 mM EGTA, 15 mM MgCl₂, 1 mM dithiothreitol, 0.1% Tween 20, 1× complete Tm protease inhibitors (Roche), 1 mM NaF, 0.2 mM NaV, 2 mM Na-pyrophosphate, 60 mM β -glycerophosphate, was homogenized for 4 × 30 s with 30 s on ice between homogenizations. Protein from supernatants (40 μg) was separated on 10% SDS-polyacrylamide (PAGE) gels and transferred to Hybond C-Super membrane (Amersham), blocked (5% non-fat milk in PBS, 0.02% Tween 20; PBST) for 1 h at room temperature and incubated with a 1/1,000 dilution of antiserum in 1% non-fat milk PBST overnight at room temperature. Washed membranes were incubated with peroxidase labelled protein A in 1% non-fat milk PBST for 1 h, washed, and developed using Amersham ECL reagents. For immunoprecipitations and kinase assays, 0.5–1 mg of extract was pre-incubated with 20 μl of protein A-Sepharose (50% suspension) for 30 min at 4 °C. The supernatant was incubated with 1 μl of antiserum (2 h on ice), 20 μl protein A-Sepharose added and the sample rotated at 4 °C for 1 h. Samples were washed 4 times with 1 ml 50 mM Tris-HCl pH 7.5, 250 mM NaCl, 5 mM EDTA, 5 mM NaF, 0.1% Tween 20, 0.5 mM PMSF and resuspended in loading buffer (for protein gel blots), or were washed twice in kinase buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 15 mM EGTA, 1 mM DTT), resuspended in 15 μl assay buffer (50 mM Tris pH 7.5, 100 mM NaCl, 5 mM EGTA, 10 mM MgCl₂, 1 mM DTT, 1 mM NaF, 0.2 mM Na-vanadate, 2 mM Na pyrophosphate, 25 mM β -glycerophosphate, 0.5 mg ml⁻¹ histone H1, 0.5 mM PMSF and 74 kBq [γ -³²P]ATP (>185 TBq mmol⁻¹) per 15 μl reaction), and incubated at room temperature for 30 min. The reaction was stopped by adding gel loading buffer, and samples were analysed by SDS-PAGE and quantitated using a phosphorimager (Molecular Dynamics). Note *in vitro* assembled plant *CycD* kinases exhibit similar levels of activity against both histone H1 and Rb protein substrates (M. Sekine, personal communication). Polyclonal rabbit antibodies were raised against full-length *CycD2;At* expressed in *Escherichia coli* and against the carboxy-terminal peptide ARNALEHEYFK-DIGYVP for *Cdc2a;Nt* (ref. 13).

Scanning electron microscopy

Seedlings fixed in 3.7% formaldehyde/0.25% glutaraldehyde in 0.1 M potassium phosphate buffer pH 7.0 at 4 °C for several hours to overnight were washed, transferred to 1%

osmium tetroxide in 0.05 M cacodylate buffer pH 7.0 at 4 °C for 2 h, rinsed in 0.025 M sodium phosphate (pH 7.0), dehydrated in a graded ethanol series at 4 °C and critical-point-dried in a Polaron E3000 Critical Point Drier using CO₂. Seedlings were mounted on stubs, and organs obscuring the apical meristem were removed. Specimens were coated with 20 nm of gold–palladium in a Polaron E5000 Sputter Coater, and examined on a Philips XL30FEG scanning electron microscope at 5 kV.

Statistical analysis

Statistical tests were performed using GraphPad Instat Version 3.0 (GraphPad Software).

Received 12 November 1999; accepted 23 March 2000.

1. Lyndon, R. F. *The Shoot Apical Meristem* (Cambridge Univ. Press, Cambridge, 1998).
2. Clark, S. E. & Schiefelbein, J. W. Expanding insights into the role of cell proliferation in plant development. *Trends Cell Biol.* **7**, 454–458 (1997).
3. Kinsman, E. A. et al. Elevated CO₂ stimulates cells to divide in grass meristems: a differential effect in two natural populations of *Dactylis glomerata*. *Plant Cell Envir.* **20**, 1309–1316 (1997).
4. Francis, D. in *Inherent Variation in Plant Growth* (eds Lambers, H. & Van Vuuren, M. M. L.) 5–20 (Backhuys, Leiden, 1998).
5. Francis, D. & Barlow, P. W. Temperature and the cell cycle. *Symp. Soc. Exp. Biol.* **42**, 181–201 (1988).
6. Hagemann, W. The relationship of anatomy to morphology in plants: A new theoretical perspective. *Int. J. Plant Sci.* **153**, 538–548 (1992).
7. Barlow, P. W. in *Shape and Form in Plants and Fungi* (eds Ingram, D. S. & Hudson, A.) 169–193 (Academic, London, 1995).
8. Soni, R., Carmichael, J. P., Shah, Z. H. & Murray, J. A. H. A family of cyclin D homologs from plants differentially controlled by growth regulators and containing the conserved retinoblastoma protein interaction motif. *Plant Cell* **7**, 85–103 (1995).
9. Bayliss, M. W. in *The Cell Division Cycle in Plants* (eds Bryant, J. A. & Francis, D.) 157–177 (Cambridge Univ. Press, Cambridge, 1985).
10. Pines, J. Cyclins and cyclin-dependent kinases: Theme and variations. *Adv. Cancer Res.* **66**, 181–212 (1995).
11. Murray, J. A. H. et al. in *Plant Cell Division* (eds Francis, D., Dudits, D. & Inzé, D.) 99–127 (Portland, London, 1998).
12. Riou-Khamlich, C., Huntley, R., Jacquard, A. & Murray, J. A. H. Cytokinin activation of *Arabidopsis* cell division through a D-type cyclin. *Science* **283**, 1541–1544 (1999).
13. Setiady, Y. Y., Sekine, M., Hariguchi, N., Kouchi, H. & Shimmyo, A. Molecular cloning and characterization of a cDNA clone that encodes a *cdc2* homolog from *Nicotiana tabacum*. *Plant Cell Physiol.* **37**, 369–376 (1996).
14. Mironov, V., De Veylder, L., Van Montagu, M. & Inzé, D. Cyclin-dependent kinases and cell division in plants—the nexus. *Plant Cell* **11**, 509–521 (1999).
15. Lyndon, R. F. Rates of cell division in the shoot apical meristem of *Pisum*. *Ann. Bot.* **34**, 1–17 (1970).
16. Galli, M. G. & Sala, F. Aphidicolin as synchronizing agent in root tip meristems of *Haploppappus gracilis*. *Plant Cell Rep.* **2**, 156–159 (1983).
17. Francis, D., Davis, M. S., Braybrook, C., James, N. C. & Herbert, R. J. An effect of zinc on M-phase and G1 of the plant cell cycle in the synchronous TBY-2 tobacco cell suspension. *J. Exp. Bot.* **46**, 1887–1894 (1995).
18. Polymenis, M. & Schmidt, E. V. Coordination of cell growth with cell division. *Curr. Opin. Genet. Dev.* **9**, 76–80 (1999).
19. Gonthier, R., Jacquard, A. & Bernier, G. Occurrence of two cell subpopulations with different cell-cycle durations in the central and peripheral zones of the vegetative shoot apex of *Sinapis alba* L. *Planta* **165**, 288–291 (1985).
20. Clowes, F. A. L. in: *Cell Division in Higher Plants* (ed. Yeoman, M. M.) 253–284 (Academic, London, 1976).
21. Gonthier, R., Jacquard, A. & Bernier, G. Changes in cell-cycle duration and growth fraction in the shoot meristem of *Sinapis* during floral transition. *Planta* **170**, 55–59 (1987).
22. Evans, H. J., Neary, G. J. & Tonkinson, S. M. The use of colchicine as an indicator of mitotic rate in broad bean root meristems. *J. Genet.* **55**, 487–502 (1957).
23. Neufeld, T. P., de la Cruz, A. F. A., Johnston, L. A. & Edgar, B. A. Coordination of growth and cell division in the *Drosophila* wing. *Cell* **93**, 1183–1193 (1998).
24. Nash, R., Tokiwa, G., Anand, S., Erickson, K. & Futcher, A. B. The *WHI1*⁺ gene of *Saccharomyces cerevisiae* tethers cell division to cell size and is a cyclin homolog. *EMBO J.* **7**, 4335–4346 (1988).
25. Quelle, D. E. et al. Overexpression of mouse D-type cyclins accelerates G₁ phase in rodent fibroblasts. *Genes Dev.* **7**, 1559–1571 (1993).
26. Resnitzky, D., Gossen, M., Bujard, H. & Reed, S. I. Acceleration of the G₁/S phase transition by expression of cyclins D1 and E with an inducible system. *Mol. Cell. Biol.* **14**, 1669–1679 (1994).
27. Doerner, P., Jorgensen, J.-E., You, R., Steppuhn, J. & Lamb, C. Control of root growth and development by cyclin expression. *Nature* **380**, 520–523 (1996).
28. Hemerly, A. et al. Dominant negative mutants of the Cdc2 kinase uncouple cell division from iterative plant development. *EMBO J.* **14**, 3925–3936 (1995).
29. Gleave, A. P. A versatile binary vector system with a T-DNA organisational structure conducive to efficient integration of cloned DNA into the plant genome. *Plant Mol. Biol.* **20**, 1203–1207 (1992).
30. Riou-Khamlich, C., Menges, M., Healy, J. M. S. & Murray, J. A. H. Sugar control of the plant cell cycle: Differential regulation of *Arabidopsis* D-type cyclin gene expression. *Mol. Cell. Biol.* (in the press).

Acknowledgements

We thank M. Sekine for permission to cite unpublished data; I. Furner for making possible the glasshouse experiments; R. Day and A. Mills for expert plant care and measurements; A. Inskip for technical support, T. Burgess and the Multi-Imaging Centre for help with scanning electron micrographs; and M. Cornelissen and M. De Block for invaluable discussions. This work was supported in part by the UK Biotechnology and Biological Sciences Research Council.

Correspondence and requests for materials should be addressed to J.A.H.M. (e-mail: j.murray@biotech.cam.ac.uk).

Peptides accelerate their uptake by activating a ubiquitin-dependent proteolytic pathway

Glenn C. Turner*, Fangyong Du* & Alexander Varshavsky

Division of Biology, California Institute of Technology, Pasadena, California 91125, USA

* These authors contributed equally to this work

Protein degradation by the ubiquitin system controls the intracellular concentrations of many regulatory proteins. A protein substrate of the ubiquitin system is conjugated to ubiquitin through the action of three enzymes, E1, E2 and E3, with the degradation signal (degron) of the substrate recognized by E3 (refs 1–3). The resulting multi-ubiquitylated substrate is degraded by the 26S proteasome⁴. Here we describe the physiological regulation of a ubiquitin-dependent pathway through allosteric modulation of its E3 activity by small compounds. Ubr1, the E3 enzyme of the N-end rule pathway (a ubiquitin-dependent proteolytic system) in *Saccharomyces cerevisiae* mediates the degradation of Cup9, a transcriptional repressor of the peptide transporter Ptr2 (ref. 5). Ubr1 also targets proteins that have destabilizing amino-terminal residues⁶. We show that the degradation of Cup9 is allosterically activated by dipeptides with destabilizing N-terminal residues. In the resulting positive feedback circuit, imported dipeptides bind to Ubr1 and accelerate the Ubr1-dependent degradation of Cup9, thereby de-repressing the expression of Ptr2 and increasing the cell's capacity to import peptides. These findings identify the physiological rationale for the targeting of Cup9 by Ubr1, and indicate that small compounds may regulate other ubiquitin-dependent pathways.

The rate of degradation of specific proteins is often regulated by modulating the exposure or the structure of their degrons. For example, the degrons of the cyclin-dependent kinase inhibitors Sic1 and p27 are activated by phosphorylation, which is timed to bring about their destruction at key transition points in the cell cycle⁷. In other cases, phosphorylation regulates the activity of an E3 itself. For example, the anaphase-promoting complex, a multisubunit E3, is activated only at mitosis⁸. One ubiquitin-dependent proteolytic system, called the N-end rule pathway, targets proteins carrying a degradation signal called the N-degron⁹, which comprises a destabilizing N-terminal residue and a lysine residue^{10,11}. In *S. cerevisiae*, there are two classes of destabilizing residues, basic or 'type 1' (Arg, Lys and His), and bulky hydrophobic or 'type 2' (Phe, Leu, Tyr, Trp and Ile). Ubr1, a RING-H2 finger-containing E3 of relative molecular mass 225,000 (M_r 225K), directly recognizes these N-terminal residues^{6,12}. A complex of Ubr1 and the E2 enzyme Rad6 (Ubc2) synthesizes a multi-ubiquitin chain linked to a lysine residue of the substrate¹³. Dipeptides with a destabilizing N-terminal residue of either basic or hydrophobic type act as competitive inhibitors of the degradation of N-end rule substrates carrying the same type of destabilizing residue^{14–16}. Thus, Ubr1 contains two distinct N-terminal residue-binding sites that are each capable of binding either a dipeptide or a protein, but not both at the same time.

The N-end rule pathway was discovered through the use of engineered reporter proteins⁹. The first physiological function of Ubr1 in *S. cerevisiae* has recently been shown to be the regulation of peptide uptake¹⁷, through control of degradation of the M_r 35K homeodomain protein Cup9, a transcriptional repressor of the di- and tripeptide transporter Ptr2 (ref. 5). Ubr1 targets Cup9 through a degron located in the carboxy-terminal half of Cup9 (F. Navarro-Garcia, G.C.T. and A.V., unpublished data). Despite this unexpected

mode of recognition, we asked whether dipeptides bearing destabilizing N-terminal residues could affect the Ubr1-mediated degradation of Cup9, as dipeptides are able to inhibit degradation of canonical N-end rule substrates¹⁶.

We addressed this question using Cup9_{NSF}, a C-terminally Flag-tagged variant of the Cup9 repressor bearing an Asn 265→Ser substitution. Cup9_{NSF} was degraded indistinguishably from wild-type Cup9 but was predicted to have a much lower affinity for DNA^{18,19}, so it would not influence the expression of Ptr2 and the uptake of dipeptides. Cup9_{NSF} was expressed as part of a fusion of the form Flag–DHFR–ubiquitin–Cup9_{NSF} (where Flag–DHFR is N-terminally Flag-tagged mouse dihydrofolate reductase; see Methods; Fig. 1). Ubiquitin-specific proteases co-translationally cleave this UPR (ubiquitin–protein–reference) fusion at the ubiquitin–Cup9 junction, yielding the test protein Cup9_{NSF} and the long-lived Flag–DHFR–ubiquitin reference protein, which serves as an internal control for variations in expression levels and immunoprecipitation efficiency^{20,21}.

Cells expressing Cup9_{NSF} were grown in minimal medium containing allantoin as the nitrogen source to avoid the known effects of nitrogen catabolite repression on *PTR2* expression²². Leu–Ala and Arg–Ala, dipeptides bearing either type of destabilizing N-terminal residue (Leu, bulky hydrophobic; Arg, basic) were added to a final concentration of 10 mM (see Methods). This dipeptide concentration results in maximal inhibition of degradation of N-end rule substrates¹⁶. Notably, the addition of either Leu–Ala or Arg–Ala exerted an opposite effect on Cup9_{NSF}, strongly accelerating its degradation in wild-type (*UBR1*) cells. The half-life of Cup9_{NSF} decreased from ~5 min in the absence of dipeptides (Fig. 1c) to less than 1 min in their presence (Fig. 1b). This stimulatory effect was not observed in a *ubr1Δ* strain, indicating that the augmented

degradation of Cup9_{NSF} was dependent on Ubr1. The enhancement of degradation required dipeptides with destabilizing N-terminal residues; dipeptides of the same composition but bearing a stabilizing residue (Ala–Leu and Ala–Arg) did not affect the degradation of Cup9_{NSF} ($t_{1/2}$ ~5 min) (Fig. 1b, and data not shown). Similar results were obtained with cells expressing Cup9_{NSF} that was not a part of a UPR fusion (data not shown).

To determine the concentration dependence of the stimulation, we measured the degradation of Cup9 at a range of concentrations of Trp–Ala. The enhancement of Cup9_{NSF} degradation was detectable at 1 μM Trp–Ala, the lowest concentration tested ($t_{1/2}$ ~1 min) (Fig. 1c). In contrast, the degradation of Cup9_{NSF} was not altered by either Ala–Trp or the constituent amino acids Trp and Ala (Fig. 1c). Similar results were obtained using Leu–Ala and Arg–Ala (data not shown). These results indicate that the relevant signalling molecule in this process is a dipeptide bearing a destabilizing N-terminal residue. The range of dipeptide concentrations that significantly stimulated Cup9 degradation was similar to physiologically active levels of many other nutrients.

Cup9 represses transcription of the transporter-encoding *PTR2*

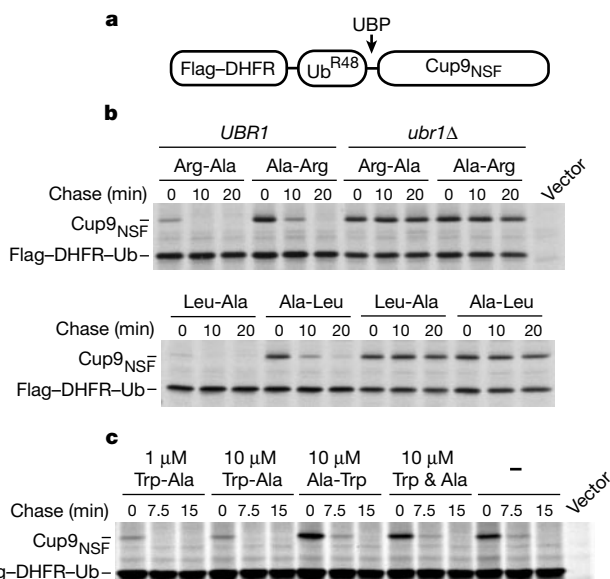


Figure 1 Enhancement of Cup9 degradation by dipeptides with destabilizing N-terminal residues. **a**, The fusion protein used for pulse-chase analysis. The stable Flag–DHFR–ubiquitin reference portion of the fusion is co-translationally cleaved from Cup9_{NSF} by ubiquitin-specific protease (UBP). **b**, Pulse-chase analysis of Flag–DHFR–ubiquitin–Cup9_{NSF} in the presence of various dipeptides at a concentration of 10 mM. Dipeptides with either basic (Arg–Ala) or bulky hydrophobic (Leu–Ala) destabilizing N-terminal residues strongly enhance Cup9_{NSF} degradation, but only in strains expressing Ubr1. Dipeptides bearing a stabilizing N-terminal residue (Ala–Arg and Ala–Leu) do not alter Cup9_{NSF} degradation. **c**, Effects of different concentrations of Trp–Ala on the enhancement of Cup9_{NSF} degradation in wild-type (*UBR1*) cells. Dash indicates that pulse-chase analysis was performed in the absence of added dipeptides. Enhancement of Cup9_{NSF} degradation was detectable at 1 μM Trp–Ala Ub^{R48}, ubiquitin containing mutation of Lys 48 to Arg (ref. 20).

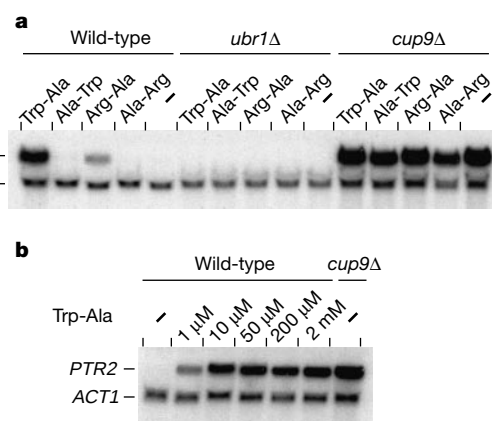


Figure 2 Effects of dipeptides on expression of the dipeptide transporter gene *PTR2*. **a**, Induction of *PTR2* expression by dipeptides bearing destabilizing N-terminal residues (Trp–Ala and Arg–Ala) required both *UBR1* and *CUP9*. Dipeptides bearing a stabilizing N-terminal residue (Ala–Trp and Ala–Arg) had no effect on *PTR2* expression. *PTR2* mRNA and the *ACT1* mRNA loading control are indicated. **b**, Effect of different concentrations of Trp–Ala on the levels of *PTR2* mRNA.

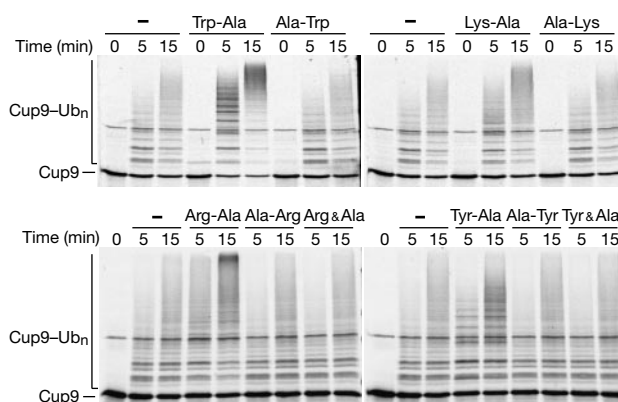


Figure 3 *In vitro* ubiquitylation of Cup9 is enhanced by dipeptides bearing destabilizing N-terminal residues. Reactions consisting of purified Uba1 (E1), Rad6 (E2), Ubr1 (E3), ubiquitin, ATP and radiolabelled Cup9 were supplemented with the indicated dipeptides or amino acids (top panel, 2 μM, bottom panel, 10 μM), or left unsupplemented (–), and allowed to proceed at 30 °C for the designated times. Radiolabelled input Cup9, and its multi-ubiquitylated derivatives are indicated. Ub, ubiquitin.

gene⁵. Thus, the dipeptide-induced acceleration of Cup9 degradation would be expected to increase the levels of *PTR2* messenger RNA, ultimately leading to an increase in dipeptide uptake. We tested this by examining the levels of *PTR2* mRNA in the presence or absence of dipeptides in the medium. At 25 μ M, both Trp–Ala and Arg–Ala induced *PTR2* expression in the wild-type (*UBR1*) strain (Fig. 2a). Both Ubr1 and Cup9 were required for these effects, as the expression of *PTR2* was not altered by dipeptides in *ubr1* Δ and *cup9* Δ strains. Induction of *PTR2* mRNA could be observed at 1 μ M Trp–Ala, and increased at higher dipeptide concentrations (Fig. 2b), in agreement with the observed changes in the half-life of Cup9 (Fig. 1c).

A plausible mechanism of the enhancement effect is that a dipeptide interacts with either the basic or the hydrophobic N-terminal residue-binding sites of Ubr1, and a distinct (third) substrate-binding site of Ubr1 recognizes the internal degenon of Cup9. In this model, the interaction of Ubr1 with dipeptides allosterically increases the ability of the Ubr1–Rad6 (E3–E2) complex to ubiquitylate Cup9. To test whether dipeptides act directly through Ubr1, we examined the effect of dipeptides on Cup9 ubiquitylation in an *in vitro* system consisting of the following purified components: Ubr1 (E3), Rad6 (E2), Uba1 (E1), ubiquitin, ATP and radiolabelled Cup9. Cup9 was significantly multi-ubiquitylated, in a Ubr1/Rad6-dependent reaction (data not shown), in

the absence of added dipeptides (Fig. 3). This result was consistent with the relatively rapid *in vivo* degradation of Cup9p ($t_{1/2}$ ~5 min) in the absence of dipeptides (Fig. 1c).

The addition of dipeptides bearing either type of destabilizing N-terminal residue to the *in vitro* system substantially stimulated the Ubr1-dependent multi-ubiquitylation of Cup9 (Fig. 3). Dipeptides of the same composition but bearing a stabilizing N-terminal residue did not stimulate multi-ubiquitylation, nor did the amino-acid components of these dipeptides (Fig. 3). These results demonstrate that dipeptides act directly through Ubr1, without an intermediate signalling pathway. The underlying allosteric mechanism may involve increased affinity of Ubr1 for Cup9, or enhanced ubiquitylation activity of the Ubr1–Rad6 complex towards Cup9, or both.

This work shows that the two binding sites of Ubr1 that interact with destabilizing N-terminal residues can act as allosteric sites that enable Ubr1 to sense the presence of imported dipeptides and to accelerate degradation of the Cup9 repressor, resulting in an appropriate induction of the *Ptr2* transporter (Fig. 4a–d). This model predicts that a dipeptide bearing a destabilizing N-terminal residue, for example Leu–Ala, should stimulate its own uptake, in contrast to Ala–Leu. This prediction was borne out when we tested the ability of these two leucine-containing dipeptides to support the growth of *S. cerevisiae* auxotrophic for leucine (Fig. 4e).

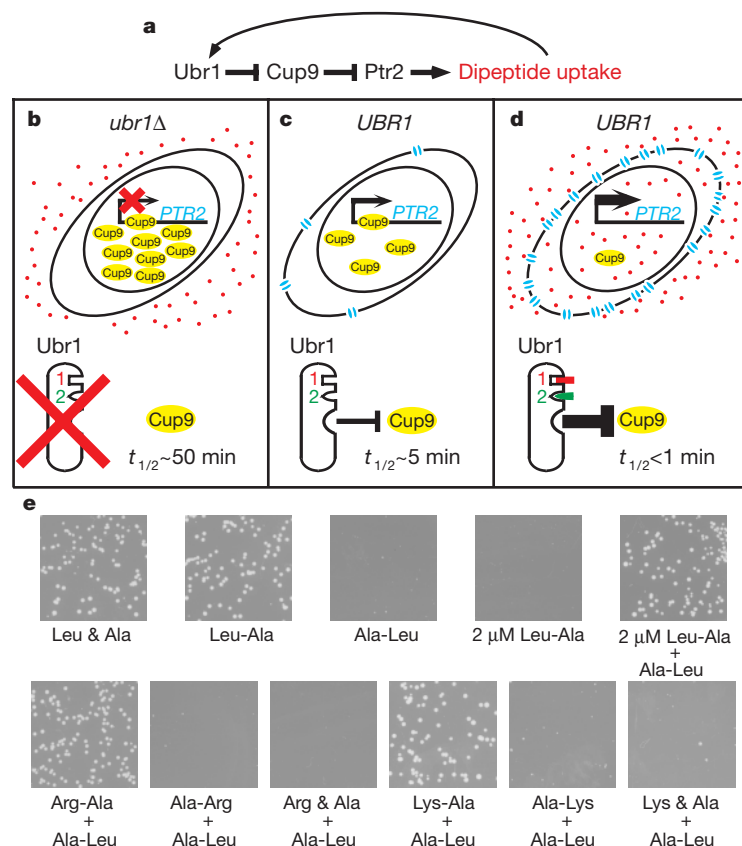


Figure 4 Feedback regulation of peptide import in *S. cerevisiae*. **a**, The peptide transport circuit. **b**, Ubr1 is required for dipeptide uptake. In the absence of Ubr1 (*ubr1* Δ), the transcriptional repressor Cup9 is long-lived, accumulates to high levels and extinguishes expression of the *PTR2* gene. Thus, *ubr1* Δ cells cannot import dipeptides (red dots). **c**, In a wild-type (*UBR1*) cell growing in the absence of extracellular dipeptides, Ubr1 targets Cup9 for degradation ($t_{1/2}$ ~5 min), resulting in a moderate concentration of Cup9 and weak but significant expression of the *Ptr2* transporter (blue double ovals). **d**, In wild-type (*UBR1*) cells growing in the presence of extracellular dipeptides, some of which bear destabilizing N-terminal residues, the imported dipeptides bind to either the basic (type 1; red) or the hydrophobic (type 2; green) residue-binding site of Ubr1. Binding of either type

of dipeptide to Ubr1 allosterically increases the rate of Ubr1-mediated degradation of Cup9. Peptides of both types are shown bound to Ubr1, but the binding of either peptide accelerates Cup9 degradation. The resulting decrease of the half-life of Cup9 from ~5 min to <1 min results in a very low concentration of Cup9, and consequently to strong induction of the *Ptr2* transporter. **e**, Colony formation assays. An *S. cerevisiae* strain requiring leucine for growth was incubated on plates supplemented with either dipeptides or their amino-acid constituents at the following concentrations: Leu–Ala, 230 μ M (2 μ M where indicated); Ala–Leu, 230 μ M; Arg–Ala, 10 μ M; Ala–Arg, 10 μ M; Lys–Ala, 1 μ M; Ala–Lys, 1 μ M; Arg & Ala, 10 μ M each; Lys & Ala, 1 μ M each.

Food sources that *S. cerevisiae* encounters outside the laboratory setting are likely to contain mixtures of short peptides, a subset of which would be capable of activating the Ubr1-based positive feedback circuit. We modelled this situation by providing cells with a mixture of Leu-Ala at a low concentration (2 μ M) and Ala-Leu at a high concentration (230 μ M). Although neither dipeptide supplement alone could satisfy the strain's requirement for leucine, a mixture of the two dipeptides supported robust growth (Fig. 4e). Moreover, a mixture of Ala-Leu (230 μ M) and a dipeptide lacking leucine but bearing a destabilizing N-terminal residue (10 μ M Arg-Ala or 1 μ M Lys-Ala) also rescued growth. In contrast, dipeptides bearing stabilizing N-terminal residues, or the amino-acid components of these dipeptides, could not rescue the growth of Leu⁻ cells in the presence of 230 μ M Ala-Leu (Fig. 4e).

This work establishes for the first time that the activity of an E3 can be directly linked to the presence of an environmental signal through an allosteric interaction with a small compound. Specifically, dipeptides bearing destabilizing N-terminal residues are shown to act as allosteric activators of Ubr1, enhancing its ability to support the ubiquitylation and degradation of Cup9. Physiologically, this results in a positive feedback circuit governing the uptake of peptides (Fig. 4). Through their binding to Ubr1, the imported dipeptides accelerate degradation of Cup9, thereby derepressing the synthesis of the Ptr2 transporter and enhancing the cell's ability to import di- and tripeptides. As most cells have the capacity to import peptides, the results of this work suggest that peptide import may be regulated similarly by Ubr1 homologues in metazoans²³ and by the ClpAP-dependent N-end rule pathway in *Escherichia coli*²⁴. The ubiquitin system is either known or suspected to be important in the control of intermediary metabolism and the transport of small molecules across membranes^{2,3}. Our findings suggest that these compounds, or their enzymatically produced derivatives, may modulate the functions of E3s in the ubiquitin system similarly to the effects observed here with dipeptides and Ubr1. □

Methods

Yeast strains and plasmids

The *S. cerevisiae* strains used in pulse-chase experiments, JD52 (*MAT α lys2-801 ura3-52 trp1- Δ 63 his3- Δ 200 leu2-3,112*), JD55 (*ubr1 Δ ::HIS3*), have been described²⁵. Flag-DHFR-ubiquitin-Cup9_{NSF} was expressed from the P_{MET25} promoter on the centromeric vector p416MET25 (ref. 26). Construction details are available upon request. Strains used for northern analyses were AVY30 (*MAT α leu2-3,112 ubr1 Δ ::LEU2*), AVY 31 (*MAT α leu2-3,112 cup9 Δ ::LEU2*) and AVY32 (*MAT α LEU2*), constructed in the RJD350 background (*MAT α leu2-3,112*; a gift from R. Deshaies) using restriction fragments obtained from plasmids pSOB30 (ref. 6), pCB119 (ref. 5) and pJJ252 (ref. 27), respectively. The RJD350 strain was used for the colony formation assay (Fig. 4e). SHM plates⁵ were supplemented with dipeptides or amino acids at the following concentrations: 230 μ M Leu-Ala (or 2 μ M where indicated), 230 μ M Ala-Leu, 230 μ M each of Leu and Ala, 10 μ M Arg-Ala, 10 μ M Ala-Arg, 1 μ M Lys-Ala, 1 μ M Ala-Lys, 10 μ M each of Arg and Ala, or 1 μ M each of Lys and Ala.

Pulse-chase analysis

Cells were cultured in SHM⁵ with auxotrophic supplements. Dipeptides were added to cultures at an absorbance of ~0.6 at 600 nm and incubation continued for 2.5 h in the experiments in Fig. 1b, and for 30 min for Fig. 1c. Cells were harvested, washed in 0.8 ml of SHM, resuspended in 0.4 ml of SHM, and labelled for 5 min at 30 °C with 0.16 mCi of ³⁵S-EXPRESS (New England Nuclear). Cells were pelleted and resuspended in fresh SHM containing 4 mM L-methionine and 2 mM L-cysteine. Samples (0.1 ml) were taken at the time points indicated and transferred to chilled tubes, each containing 0.5 ml of 0.5-mm glass beads, 0.7 ml of ice-cold lysis buffer (1% Triton-X100, 0.15 M NaCl, 5 mM EDTA, 50 mM sodium HEPES, pH 7.5), and a mixture of protease inhibitors (final concentrations 1 mM phenylmethylsulphonyl fluoride, 2 μ g ml⁻¹ aprotinin, 0.5 μ g ml⁻¹ leupeptin and 0.7 μ g ml⁻¹ pepstatin). Extracts were prepared and immunoprecipitations carried out as described²⁵, using anti-Flag M2 resin (Sigma). Immunoprecipitates were fractionated by 13% SDS-polyacrylamide gel electrophoresis (PAGE), and detected by autoradiography.

RNA preparation and northern analysis

Cells were cultured in SHM to an absorbance of ~0.6 at 600 nm. The indicated dipeptides were then added to a 25 μ M final concentration, and the incubation was continued for an

additional 30 min. Total RNA was prepared²⁸ and 25 μ g samples were electrophoresed in 1% formaldehyde-agarose gels, followed by blotting for northern analysis²⁹.

Ubr1-dependent *in vitro* ubiquitylation system

The components of this system were purified as follows. N-terminally hexahistidine-tagged Uba1 was overexpressed in *S. cerevisiae* and purified by fractionation over Ni-NTA, ubiquitin affinity and Superdex-200 columns. Rad6 was overexpressed in *E. coli* and purified by fractionation over DEAE, Mono-Q and Superdex-75 columns. N-terminally Flag-tagged Ubr1 was overexpressed in *S. cerevisiae*, and purified by fractionation over anti-Flag M2, Ubc2p affinity and Superdex-200 columns. N-terminally Flag-tagged, C-terminally hexahistidine-tagged Cup9 was expressed and radiolabelled in *E. coli*, and purified by fractionation over Ni-NTA and anti-Flag M2 columns. Details of these expression and purification protocols are available upon request.

The *in vitro* ubiquitylation reactions contained the following components: 7 μ M ubiquitin, 50 nM Uba1, 50 nM Rad6, 50 nM Ubr1, 550 nM ³⁵S-labelled Cup9, 25 mM HEPES/KOH (pH 7.5), 25 mM KCl, 5 mM MgCl₂, 2 mM ATP, 0.1 mM dithiothreitol and 0.5 mg ml⁻¹ ovalbumin as a carrier protein. Dipeptides or amino acids, as indicated, were added to a final concentration of 2 μ M (top panel, Fig. 3) or 10 μ M (bottom panel, Fig. 3). All components except Uba1 were mixed on ice for 10 min; Uba1 was then added and reactions shifted to 30 °C. After the indicated times, the reactions were terminated by adding an equal volume of two times SDS-PAGE loading buffer and heating at 95 °C for 5 min, followed by 8% SDS-PAGE.

Received 5 January; accepted 5 April 2000.

1. Laney, J. D. & Hochstrasser, M. Substrate targeting in the ubiquitin system. *Cell* **97**, 427–430 (1999).
2. Herskko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **76**, 425–479 (1998).
3. Varshavsky, A. The ubiquitin system. *Trends Biochem. Sci.* **22**, 383–387 (1997).
4. Baumeister, W., Walz, J., Zühl, F. & Seemüller, E. The proteasome: paradigm of a self-compartmentalizing protease. *Cell* **92**, 367–380 (1998).
5. Byrd, C., Turner, G. C. & Varshavsky, A. The N-end rule pathway controls the import of peptides through degradation of a transcriptional repressor. *EMBO J.* **17**, 269–277 (1998).
6. Bartel, B., Wünnig, I. & Varshavsky, A. The recognition component of the N-end rule pathway. *EMBO J.* **9**, 3179–3189 (1990).
7. King, R. W., Deshaies, R. J., Peters, J. M. & Kirschner, M. W. How proteolysis drives the cell cycle. *Science* **274**, 1652–1659 (1996).
8. Kotani, S., Tanaka, H., Yasuda, H. & Todokoro, K. Regulation of APC activity by phosphorylation and regulatory factors. *J. Cell Biol.* **146**, 791–800 (1999).
9. Bachmair, A., Finley, D. & Varshavsky, A. *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science* **254**, 179–186 (1996).
10. Bachmair, A. & Varshavsky, A. The degradation signal in a short-lived protein. *Cell* **56**, 1019–1032 (1989).
11. Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc. Natl Acad. Sci. USA* **93**, 12142–12149 (1996).
12. Xie, Y. M. & Varshavsky, A. The E2-E3 interaction in the N-end rule pathway: the RING-H2 finger of E3 is required for the synthesis of multiubiquitin chain. *EMBO J.* **18**, 6832–6844 (1999).
13. Dohmen, R. J., Madura, K., Bartel, B. & Varshavsky, A. The N-end rule is mediated by the UBC2(RAD6) ubiquitin-conjugating enzyme. *Proc. Natl Acad. Sci. USA* **88**, 7351–7355 (1991).
14. Reiss, Y., Kaim, D. & Herskko, A. Specificity of binding of N-terminal residues of proteins to ubiquitin-protein ligase. Use of amino acid derivatives to characterize specific binding sites. *J. Biol. Chem.* **263**, 2693–2699 (1988).
15. Gonda, D. K. *et al.* Universality and structure of the N-end rule. *J. Biol. Chem.* **264**, 16700–16712 (1989).
16. Baker, R. T. & Varshavsky, A. Inhibition of the N-end rule pathway in living cells. *Proc. Natl Acad. Sci. USA* **87**, 2374–2378 (1991).
17. Alagaram, K., Naider, F. & Becker, J. M. A recognition component of the ubiquitin system is required for peptide transport in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **15**, 225–234 (1995).
18. Knight, S. A. B., Tamai, K. T., Kosman, D. J. & Thiele, D. J. Identification and analysis of a *Saccharomyces cerevisiae* copper homeostasis gene encoding a homeodomain protein. *Mol. Cell. Biol.* **14**, 7792–7804 (1994).
19. Wolberger, C., Vershon, A. K., Liu, B. S., Johnson, A. D. & Pabo, C. O. Crystal structure of a Mata-2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**, 517–528 (1991).
20. Lévy, F., Johnsson, N., Rumenapf, T. & Varshavsky, A. Using ubiquitin to follow the metabolic fate of a protein. *Proc. Natl Acad. Sci. USA* **93**, 4907–4912 (1996).
21. Suzuki, T. & Varshavsky, A. Degradation signals in the lysine-asparagine sequence space. *EMBO J.* **18**, 6017–6026 (1999).
22. Barnes, D., Lai, W., Breslav, M., Naider, F. & Becker, J. M. PTR3, a novel gene mediating amino acid-inducible regulation of peptide transport in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **29**, 297–310 (1998).
23. Kwon, Y. T. *et al.* The mouse and human genes encoding the recognition component of the N-end rule pathway. *Proc. Natl Acad. Sci. USA* **95**, 7898–7903 (1998).
24. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria. *Science* **254**, 1374–1377 (1991).
25. Ghislain, M., Dohmen, R. J., Levy, F. & Varshavsky, A. Cdc48p interacts with Ufd3p, a WD repeat protein required for ubiquitin-mediated proteolysis in *Saccharomyces cerevisiae*. *EMBO J.* **15**, 4884–4899 (1996).
26. Mumberg, D., Muller, R. & Funk, M. Regulatable promoters of *Saccharomyces cerevisiae*—comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res.* **22**, 5767–5768 (1994).
27. Jones, J. S. & Prakash, L. Yeast *Saccharomyces cerevisiae* selectable markers in pUC18 polylinkers. *Yeast* **6**, 363–366 (1990).
28. Schmitt, M. E., Brown, T. A. & Trumpower, B. L. A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **18**, 3091–3092 (1990).
29. Ausubel, F. M. *et al.* (eds) *Current Protocols in Molecular Biology*. (Wiley-Interscience, New York, 1996).

Acknowledgements

We thank A. Webster for her valuable contributions to establishing the *in vitro* system, and R. Deshaies, J. Dohmen, L. Prakash, H. Rao and J. Sheng for their gifts of plasmids and strains. We also thank C. Byrd, H. Rao, T. Iverson and especially R. Deshaies for helpful discussions and comments on the manuscript. This work was supported by grants to A.V. from the NIH.

Correspondence and requests for materials should be addressed to A.V.
(e-mail: avarsh@caltech.edu).

Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule

Craig H. Bassing*, Frederick W. Alt*, Maureen M. Hughes†, Margaux D'Auteuil*, Tara D. Wehrly†, Barbara B. Woodman*, Frank Gärtner*, J. Michael White†, Laurie Davidson* & Barry P. Sleckman†

* Howard Hughes Medical Institute, Children's Hospital and Department of Genetics, Harvard Medical School and The Center for Blood Research, Boston, Massachusetts 02115, USA

† Washington University School of Medicine, Department of Pathology and Immunology, 660 S. Euclid Ave., Campus Box 8118, St. Louis, Missouri 63110, USA

The genes encoding the variable regions of lymphocyte antigen receptors are assembled from variable (V), diversity (D) and joining (J) gene segments¹. V(D)J recombination is initiated by the recombinase activating gene (RAG)-1 and -2 proteins, which introduce DNA double-strand breaks between the V, D and J segments and their flanking recombination signal sequences (RSSs). Generally expressed DNA repair proteins then carry out the joining reaction^{2,3}. The conserved heptamer and nonamer sequences of the RSSs are separated by non-conserved spacers of 12 or 23 base pairs (forming 12-RSSs and 23-RSSs). The 12/23 rule, which is mediated at the level of RAG-1/2 recognition and cutting^{4,5}, specifies that V(D)J recombination occurs only between a gene segment flanked by a 12-RSS and one flanked by a 23-RSS¹. V β segments are appended to DJ β rearrangements, with little or no direct V β to J β joining, despite 12/23 compatibility of V β 23-RSSs and J β 12-RSSs^{6,7}. Here we use embryonic stem cells and mice with a modified T-cell receptor (TCR) β locus containing only one D β (D β 1) gene segment and one J β (J β 1) gene cluster to show that the 5' D β 12-RSS, but not the J β 12-RSSs, targets rearrangement of a diverse V β repertoire. This targeting is precise and position-independent. This additional restriction on V(D)J recombination has important implications for the regulation of variable region gene assembly and repertoire development.

Assembly of TCR β genes occurs through an ordered, two-step process with D β -to-J β rearrangements occurring on both alleles before V β rearrangement. V β segments are flanked with 23-RSSs, D β segments with 5' 12-RSSs and 3' 23-RSSs, and J β segments with 12-RSSs⁸. Thus, direct V β -to-J β rearrangement, without a D β , is permissible based on the 12/23 rule. Most TCR β gene rearrangements that involve a V β also involve a D β , however, indicating that additional restrictions prevent direct V β -to-J β joining. Transgenic V(D)J recombination substrate studies have confirmed this restriction and raised the possibility that the 5' D β 12-RSSs, and not the J β 12-RSSs, may specifically target V β rearrangements^{7,31}. Recombination substrate studies have also shown that V(D)J recombination efficiency can be influenced by both RSSs and their flanking coding

and non-coding sequences^{9–18}. However, these substrate studies—which tested a limited number of sequences—may not reflect general principles of endogenous antigen receptor loci that contain numerous V, D and J gene segments with many different RSS variations.

The endogenous murine TCR β locus contains about 35 V β segments 5' of two D-J-C β clusters (Fig. 1a); each D-J-C β cluster (D β 1-J β 1-C β 1 and D β 2-J β 2-C β 2) contains a single D β (D β 1 and D β 2), six functional J β s (J β 1.1–1.6 and J β 2.1–2.7, including one unfunctional J β 2) and a single C β gene (C β 1 and C β 2). To facilitate the analysis of elements that control rearrangement, we generated an embryonic stem (ES) cell line homozygous for a mutation in which the D-J β 2 gene cluster was replaced with a single *loxP* site (J β 1^{lox} ES cell; Fig. 1a). Therefore, J β 1^{lox} allele rearrangements that involve D β must use D β 1. We used J β 1^{lox} cells to generate chimaeric mice by RAG-2 deficient blastocyst complementation; these chimaeras were also bred to introduce the J β 1^{lox} allele into the germline¹⁹. Analysis of J β 1^{lox} mice showed that thymocyte development, splenic T-cell numbers, V β utilization and J β 1 gene segment utilization were similar to those of wild-type mice (Figs 1b and 2, and data not shown). Additional analysis of J β 1^{lox} $\alpha\beta$ splenic T-cell hybridomas showed that all of them had D-J β 1 gene cluster rearrangements on both alleles (Table 1). Therefore, deletion of the D-J β 2 cluster did not appreciably alter D-J β 1 cluster rearrangements or $\alpha\beta$ T-cell development.

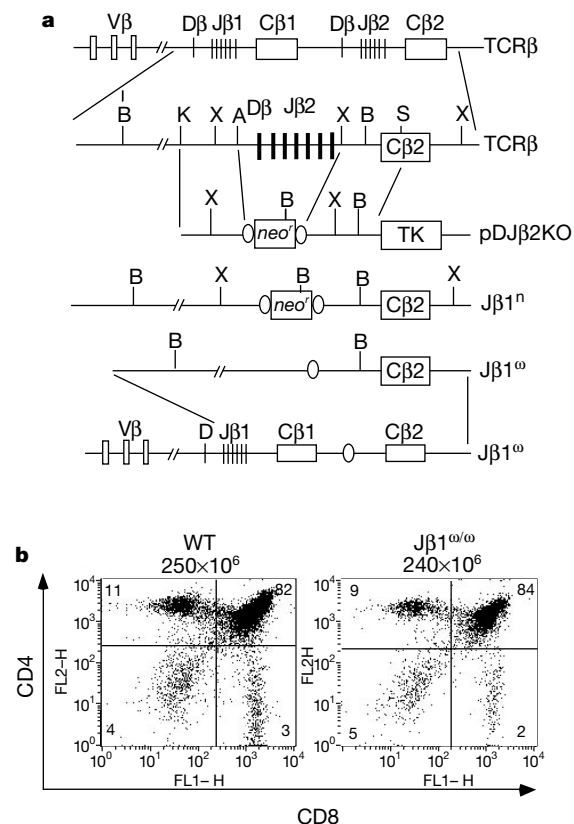


Figure 1 Generation of J β 1^{lox} ES cells. **a**, Schematic of the TCR β locus showing the DJ β 1 and DJ β 2 gene clusters with associated C β 1 and C β 2 genes and some V β gene segments (not to scale). The pDJ β 2KO vector was used to target and replace the DJ β 2 gene cluster with a neomycin resistance (*neo*) gene flanked by *loxP* sites (ovals) generating the J β 1^{lox} allele. Cre-mediated deletion of the *neo* gene left a single *loxP* site in place of the DJ β 2 gene cluster (J β 1^{lox}). *Bam*HI (B), *Kpn*I (K), *Alu*NI (A), *Xba*I (X) and *Spe*I (S) sites are indicated. **b**, Flow cytometric analysis of thymocytes from 3–4 week-old wild-type (WT) and J β 1^{lox} mice using CD4-PE and CD8-FITC. At least eight mice were analysed in each group.

To assay for specific RSS functions *in vivo*, we first generated the M4 mutation of the $J\beta 1^{\omega}$ allele, in which the 5' D $\beta 1$ 12-RSS was replaced by the $J\beta 1.2$ 12-RSS without altering any flanking sequences ($J\beta 1^{M4}$, Fig. 3a). In addition, the $J\beta 1^{M4}$ allele contained *loxP* and *HindIII* sites between D $\beta 1$ and $J\beta 1.1$, as well as the conversion of a *Bam*HI site 3' of $J\beta 1.2$ to a *Pst*I site, to distinguish between $J\beta 1^{\omega}$ and $J\beta 1^{M4}$ allele rearrangements (Fig. 3a). TCR β D-to-J rearrangement begins in CD4⁺/CD8⁺ (double negative, DN) thymocytes at the CD25⁺ stage²⁰. To assess the effects of the M4 mutation, we sorted $J\beta 1^{M4/\omega}$ and $J\beta 1^{\omega/\omega}$ CD25⁺ DN thymocytes, amplified D $\beta 1$ to $J\beta 1.1$ or $J\beta 1.2$ rearrangements by polymerase chain reaction (PCR), and distinguished between products from the $J\beta 1^{\omega}$ versus $J\beta 1^{M4}$ alleles by *Bam*HI and *Pst*I digestion, respectively; similar levels of rearrangement occurred on both alleles (Fig. 3b). Next, we assayed for $J\beta 1$ rearrangements in $J\beta 1^{M4/\omega}$ splenic T-cell hybridomas by Southern blotting and found rearrangements of both the $J\beta 1^{\omega}$ and $J\beta 1^{M4}$ alleles in all cells analysed (Fig. 3c, Table 1). However, additional Southern blot and PCR analyses showed that the $J\beta 1^{M4}$ allele lacked V β rearrangements and was in the DJ β configuration (Fig. 3d, Table 1). Therefore, neither the developmental onset nor the efficiency of D β -to- $J\beta$ rearrangement was altered on the $J\beta 1^{M4}$ allele. However, V β -to-D $\beta 1$ rearrangement on this allele is markedly diminished. Correspondingly, $J\beta 1^{M4/M4}$ mice exhibited a significant blockade in thymocyte development at the DN stage and reduced numbers of peripheral T cells (Fig. 3e, data not shown). Thus, the 5' D $\beta 1$ 12-RSS is needed to target efficient V β rearrangement beyond simply enforcing the 12/23 rule.

To assess whether D $\beta 1$ and its flanking RSSs inhibit direct joining of V β to $J\beta$, we introduced a mutation (M3) that deleted D $\beta 1$ and its flanking RSSs from the $J\beta 1^{\omega}$ allele ($J\beta 1^{M3}$, Fig. 4a). Southern blot analysis of $J\beta 1^{M3/\omega}$ $\alpha\beta$ T-cell hybridomas showed that all of them

Table 1 Analysis of TCR β rearrangements in $\alpha\beta$ T-cell hybridomas

	Number	Allele	GL	DJ	V(D)J	Other
$J\beta 1^{\omega/\omega}$	50	ω	0	34	16	0
		ω	0	0	50	0
$J\beta 1^{M4/\omega}$	69	ω	0	0	69	0
		M4	0	66	0	3*
$J\beta 1^{M3/\omega}$	75	ω	0	0	75	0
		M3	75	0	0	0
$J\beta 1^{M5/\omega}$	73	ω	0	13	60	0
		M5	51	0	22†	0

T-cell hybridomas were generated, $\alpha\beta$ TCR expression confirmed and clonality determined as described²⁰. TCR β allele configuration in $J\beta 1^{M4/\omega}$ hybridomas was determined by Southern blot and PCR analysis as described in Fig. 3. TCR β allele configuration in $J\beta 1^{\omega/\omega}$, $J\beta 1^{M3/\omega}$ and $J\beta 1^{M5/\omega}$ hybridomas was determined by Southern blot analysis of *Eco*RI-digested genomic DNA using probes A and C (Fig. 4a). GL, germline.

* $J\beta 1^{M4}$ allele rearrangements identified by cloning and sequence analyses. Two of these alleles underwent distinct D $\beta 1$ rearrangements to a cryptic heptamer that lies 7 kb downstream of the C $\beta 2$ gene (data not shown). The third $J\beta 1^{M4}$ allele has a V $\beta 5.2$ RSS to $J\beta 1.1$ hybrid join and an inverted D $\beta 1$ to V $\beta 5.2$ coding join (data not shown).

† All V $\beta 1.2$ rearrangements as determined by PCR using primers PV and P2 followed by digestion of the PCR products with *Pst*I (Fig. 4a and data not shown).

had V β (D) $J\beta 1$ rearrangements on the $J\beta 1^{\omega}$ allele and none had the $J\beta 1^{M3}$ allele rearranged (Table 1). In addition, PCR analysis demonstrated that little, if any, V β -to- $J\beta$ rearrangement occurred on the $J\beta 1^{M3}$ allele in $J\beta 1^{M3/\omega}$ thymocytes (Fig. 4b).

The 5' D $\beta 1$ 12-RSS contains a functional TATA box used by the PD β promoter for initiation of DJ $\beta 1$ germline transcripts, a process that may be important for D β -to- $J\beta$ rearrangement^{21–23}. However, northern blot analysis of RNA from DN thymocytes of $J\beta 1^{M3/M3}$; RAG-2^{−/−} mice revealed normal levels of $J\beta 1$ -hybridizing germline transcripts (Fig. 4c). We conclude that V β -to- $J\beta 1$ rearrangement fails to occur efficiently on the $J\beta 1^{M3}$ allele despite 12/23 compatibility

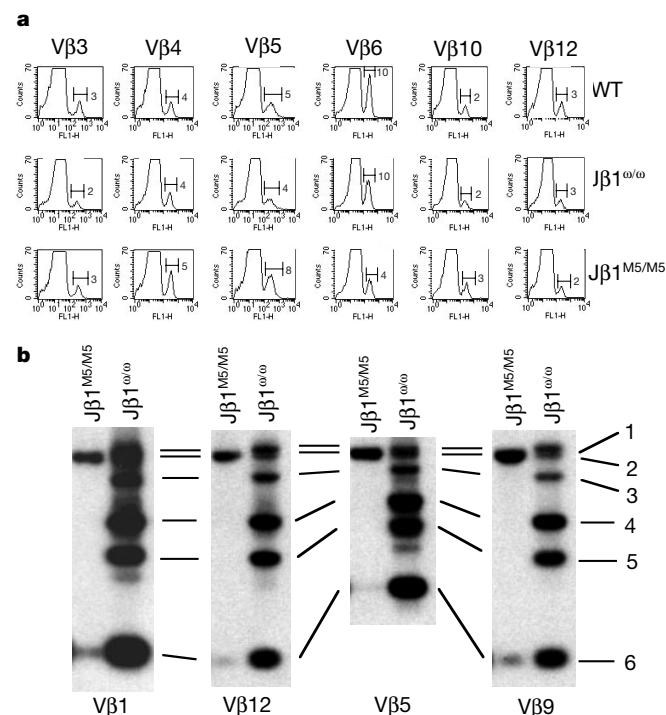


Figure 2 V β and $J\beta$ repertoire in WT, $J\beta 1^{\omega/\omega}$ and $J\beta 1^{M5/M5}$ mice. **a**, Lymph node cells were stained with Thy1-PE and biotin-conjugated anti-V $\beta 4$, V $\beta 5$, V $\beta 6$, V $\beta 8$ and V $\beta 10$ followed by streptavidin-FITC. Histograms shown are gated on Thy1⁺ cells and the percentage of V β ⁺/Thy1⁺ cells is indicated. **b**, PCR analysis of $J\beta 1^{\omega/\omega}$ and $J\beta 1^{M5/M5}$ thymocyte DNA using the PV primer set and the P4 primer (Fig. 4a). PCR products were analysed by Southern blot using the PR3 oligonucleotide probe (Fig. 4a). Expected sizes for PCR products for V β J1.1 through V β J1.6 rearrangements (1 through 6) are indicated.

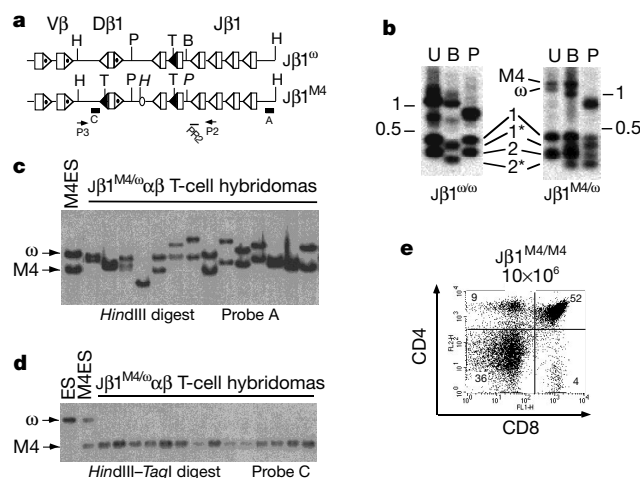


Figure 3 Rearrangement of the $J\beta 1^{M4}$ allele. **a**, Schematic comparison of the $J\beta 1^{M4}$ and $J\beta 1^{\omega}$ alleles. The position of the $J\beta 1.2$ 12-RSS (shaded triangle, 5'-CATATTCGAGTATCTGATTCTGATGTG-3') with its native *Taq*I site (T) is shown. Native 23-RSSs (dotted open triangles) and 12-RSSs (open triangles) are indicated. The *loxP* site (oval), native (H) and introduced (H) *Hind*III sites, and the *Bam*HI (B) to *Pst*I (P) conversion, the locations of probes A and C, and the P3, P2 and PR2 oligonucleotides are shown. **b**, Southern blot analysis of undigested (U), *Bam*HI (B)- or *Pst*I (P)- digested PCR products from CD25⁺ DN thymocytes. PCR products from D $\beta 1$ to $J\beta 1.1$ (1) or $J\beta 1.2$ (2) rearrangements and the product size after digestion (1* and 2*) with *Bam*HI ($J\beta 1^{\omega}$) or *Pst*I ($J\beta 1^{M4}$) are indicated. The undigested products from PCR of germline $J\beta 1^{M4}$ (M4) and $J\beta 1^{\omega}$ (ω) alleles, and the 1 and 0.5-kb markers are shown. **c**, **d**, Southern blot analysis of *Hind*III (**c**) and *Hind*III–*Taq*I (**d**)-digested genomic DNA from $J\beta 1^{M4/\omega}$ T-cell hybridomas and wild-type (ES) and $J\beta 1^{M4/\omega}$ (M4ES) ES cells with probes A (**c**) and C (**d**). Bands from unrearranged $J\beta 1^{M4}$ (M4) and $J\beta 1^{\omega}$ (ω) alleles are indicated. **e**, Flow cytometric analysis of thymocytes from 3–4 week-old $J\beta 1^{M4/M4}$ mice as described in the legend to Fig. 1.

and germline transcription of the J β 1 region. Consistent with this finding, J β 1^{M3/M3} mice exhibit a blockade of thymocyte development and diminution of peripheral T-cell numbers similar in severity to that observed in J β 1^{M4/M4} mice (Fig. 4d). The few mature T cells generated in J β 1^{M3/M3} mice are due to direct V β -to-J β 1 rearrangements that employ a diverse set of V β segments and all of the J β 1 segments (data not shown). These findings demonstrate that efficient joining of V β and J β segments on the J β 1^ω allele must be mediated by intermediate rearrangements involving D β 1.

To assess further the role of the 5' D β 1 RSS in targeting V β rearrangements, we generated the J β 1^{M5} mutant allele. This allele is identical to J β 1^{M3} (deletion of D β 1) but also has a mutation in which the J β 1.2 12-RSS was replaced with the 5' D β 1 12-RSS (Fig. 4a). Southern blot analysis of 71 J β 1^{M5/ω} α β T-cell hybridomas revealed that 22 had rearrangements of their J β 1^{M5} allele (Table 1). Furthermore, the J β 1^ω allele was in the DJ β 1 configuration in 13 hybridomas, indicating that the J β 1^{M5} allele had undergone a functional V β -to-J β rearrangement that is expressed by these cells

(Table 1). Additional analysis confirmed that all 22 of the rearranged J β 1^{M5} alleles had V β -to-J β 1.2 rearrangements (data not shown). PCR analysis of J β 1^{M5/ω} thymocyte DNA revealed no detectable VJ β 1.1 rearrangements on the J β 1^{M5} allele, whereas about half of the VJ β 1.2 rearrangements occurred on the J β 1^{M5} allele (Fig. 4b). J β 1^{M5/M5} mice exhibited essentially normal thymocyte development and accumulation of peripheral T cells expressing TCR β chains that utilize a diverse repertoire of V β gene segments (Figs 2a and 4c, and data not shown). However, V β rearrangements in J β 1^{M5/M5} thymocytes were predominantly to the J β 1.2 gene segment (Fig. 2b).

Our findings demonstrate that the 5' D β 1 12-RSS targets rearrangement of diverse endogenous V β s through their 23-RSSs. Furthermore, because the J β 1 12-RSSs do not support V β 23-RSS-mediated rearrangements but do support 3' D β 1 23-RSS-mediated rearrangements, the 5' D β 1 12-RSS must fulfil its role in targeting V β s based on properties beyond those of simple 12/23 compatibility. This specific recombination targeting process is very precise. The 5' D β 1 12-RSS transplanted to J β 1.2 on the J β 1^{M5} allele targets V β gene segments from as far as 500 kilobases (kb) upstream specifically to the J β 1.2 segment. Moreover, this specific targeting occurs despite the presence of upstream and downstream J β 12-RSSs in very close proximity (approximately 100 base pairs (bp)). In addition, the function of the 5' D β 1 12-RSS is position-independent. Moving this RSS 1 kb downstream from its normal location targets V β rearrangement to the new location (J β 1^{M5} allele).

To account for this restriction in the endogenous V(D)J recombination process, the mechanism responsible must be based on nucleotide sequence differences between the 5' D β 1 12-RSS and the various J β 1 12-RSSs. The presence of a functional TATA box in the 5' D β 1 RSS raises the possibility that this sequence could promote V β segment rearrangement by transcription-related mechanisms. However, such transcriptional targeting would have to be linked to transcription initiation, as J β segments on the J β 1^{M3} allele are not efficient substrates for V β rearrangement, despite their continued germline transcription. There are six evolutionarily conserved nucleotides in the spacers of 5' D β RSSs, a level of conservation not observed in J β RSS spacers²⁴. It is conceivable that these or other less obvious differences in RSS sequences might provide *cis*-acting constraints that permit functional synapses between V β /5' D β RSS and 3' D β /J β RSS combinations but not V β /J β RSS combinations. A final possibility would be the existence of developmental-stage or lineage-specific factors that alter the ability of RAG-1/2 to use specific RSSs as substrates. In any case, the marked specificity of RSS utilization could have many implications for the generation of specific V(D)J repertoires. For example, this phenomenon could, at least in part, explain differential rearrangement of V gene segments within a given locus²⁵ or the specific sorting of interspersed V α and V δ gene families²⁶.

Ordered D-to-J followed by V-to-DJ rearrangements was first observed in the IgH locus²⁷. As V_H and J_H segments have 23-RSSs and D_H segments have 5' and 3' 12-RSSs, direct V_H-to-J_H rearrangements are prohibited by the 12/23 rule, ensuring D_H segment utilization in IgH variable region genes. Although this RSS organization prevents direct V_H-to-J_H joining, it does not specify a sequential order of assembly steps. Our finding that V β segments do not recombine efficiently with J β segments despite 12/23 compatibility emphasizes the importance of D β utilization, possibly to ensure CDR3 diversification and length²⁸. However, this novel restriction mechanism cannot fully explain preferential DJ β assembly before appendage of a V β to a D β . Although the precise mechanisms responsible for this ordered rearrangement process are not known, we have found that D β to J β rearrangement *per se* is not required for V β -to-D β rearrangement³¹. Moreover, order appears to be required for feedback regulation of the IgH and TCR β V-to-DJ rearrangement step, a process mediated at the level of accessibility to the RAG proteins^{6,27}. Our findings raise the

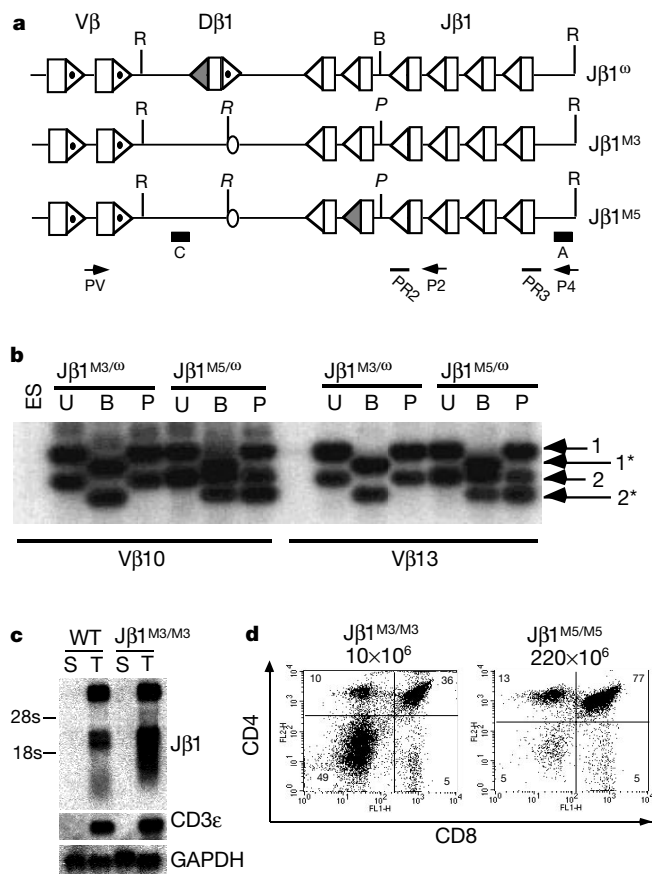


Figure 4 Rearrangement analysis of J β 1^{M3} and J β 1^{M5} alleles. **a**, The J β 1^ω, J β 1^{M3} and J β 1^{M5} alleles. The native (R) and introduced (R) *Eco*RI sites are shown and the position of the 5' D β 1 12-RSS (shaded triangle, 5'-CTTTTGTGATAAAGCTGTACATTGTG-3') is indicated. The locations of the PR3 and P4 primers, and PV primer set are shown. All other details are as in Fig. 3a. **b**, Southern blot analysis was carried out on undigested (U), *Bam*HI-(B) and *Pst*I-(P) digested PCR products of PV/P2 amplified genomic DNA from J β 1^{M3/ω} and J β 1^{M5/ω} thymocytes. Products from V β to J β 1.1 (1) and V β to J β 1.2 (2) rearrangements and the expected sizes of the *Bam*HI or *Pst*I-digested bands (1* and 2*) are indicated. Analyses with V β 1, 2, 3, 5, 8, 11 and 14 primers yielded similar results and are not shown. **c**, RNA was isolated from thymocytes (T) and splenocytes (S) of RAG-2^{-/-} (WT) or J β 1^{M3/M3};RAG-2^{-/-} (J β 1^{M3/M3}) mice and subjected to northern blot analysis using a probe that spanned the J β 1 gene segments (J β 1) and the CD3 ϵ and GAPDH probes as controls. **d**, Flow cytometric analysis of thymocytes from 3–4-week-old J β 1^{M3/M3} and J β 1^{M5/M5} mice as described in the legend to Fig. 1.

possibility that, at least in some contexts, accessibility regulation could involve the modulation of putative RSS-specific *trans*-acting factors that promote RAG-1/2 access. □

Methods

Generation of targeted ES cells and mutant mice

We generated pDJ β 2KO by subcloning the 4-kb *KpnI*–*AlwNI* 5' and 3-kb *XbaI*–*SpeI* 3' homology regions that flank the DJ β 2 gene cluster into the previously described pLNTK targeting vector²⁹. The construction of the targeting vectors for the M3, M4 and M5 mutations and the generation of ES cells with J β 1^{M3}, J β 1^{M4} and J β 1^{M5} mutant alleles are described in the Supplementary Information. Transfection and selection of ES cells were carried out as described²⁹. Chimaeric mice were generated by RAG-2-deficient blastocyst complementation¹⁹.

Southern and northern blot analyses

Southern and northern blot analyses were carried out as described²⁹. Probes A and C were 0.8-kb *DrdI*–*DrdI* and 0.7-kb *AflII*–*HaeII* TCR β genomic DNA fragments, respectively. The J β 1 probe was a 2-kb *PstI* genomic DNA fragment. Glyceraldehyde phosphate dehydrogenase and complementary CD3 ϵ DNA probes have been described previously²⁹.

PCR analyses

The V β primer set (PV) has been described previously³⁰ with the exception of the V β 14 primer: 5'-GGCAAGCAAGCTGGTGTGT-3'. The sequences of the remaining primers are as follows: P2, 5'-CCTGACTTCCACCCGAGGTT-3'; P3, 5'-CCTTCCTTATCTT-CAACTC-3'; P4, 5'-AAGGACGACTCTGTCTT-3'; PR2, 5'-CATCCTTCCTCTGAT-TAC-3'. PCR was carried out as described³⁰ at 92°, 90 s; 60°, 150 s; 72°, 60 s for 30 cycles.

Flow cytometric analysis and cell sorting

Flow cytometric analysis of thymocytes for CD4 and CD8 expression was carried out as described²⁹. Cell sorting was carried out by pretreatment of thymocytes with anti-CD4 and anti-CD8 and complement followed by sorting of CD25⁺/CD4⁺/CD8⁺ thymocytes on an Epics MoFlow sorter.

Received 8 March; accepted 19 April 2000.

1. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
2. Weaver, D. T. V(D)J recombination and double-strand break repair. *Adv. Immunol.* **58**, 29–85 (1995).
3. Jeggo, P. A. Identification of genes involved in repair of DNA double-strand breaks in mammalian cells. *Radiat. Res.* **150**, 580–591 (1998).
4. Eastman, Q. M., Leu, T. M. & Schatz, D. G. Initiation of V(D)J recombination *in vitro* obeying the 12/23 rule. *Nature* **380**, 85–88 (1996).
5. van Gent, D. C., Ramsden, D. A. & Gellert, M. The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. *Cell* **85**, 107–113 (1996).
6. Uematsu, Y. *et al.* In transgenic mice the introduced functional T cell receptor beta gene prevents expression of endogenous beta genes. *Cell* **52**, 831–841 (1988).
7. Ferrier, P. *et al.* Separate elements control DJ and VDJ rearrangement in a transgenic recombination substrate. *EMBO J.* **9**, 117–125 (1990).
8. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
9. Nadel, B., Tang, A., Escuro, G., Lugo, G. & Feeney, A. J. Sequence of the spacer in the recombination signal sequence affects V(D)J rearrangement frequency and correlates with nonrandom V κ usage *in vivo*. *J. Exp. Med.* **187**, 1495–1503 (1998).
10. Hesse, J. E., Lieber, M. R., Mizuuchi, K. & Gellert, M. V(D)J recombination: a functional definition of the joining signals. *Genes Dev.* **3**, 1053–1061 (1989).
11. Gerstein, R. M. & Lieber, M. R. Coding end sequence can markedly affect the initiation of V(D)J recombination. *Genes Dev.* **7**, 1459–1469 (1993).
12. Boubnov, N. V., Wills, Z. P. & Weaver, D. T. Coding sequence composition flanking either signal element alters V(D)J recombination efficiency. *Nucleic Acids Res.* **23**, 1060–1067 (1995).
13. Ramsden, D. A. & Wu, G. E. Mouse κ light-chain recombination signal sequences mediate recombination more frequently than do those of λ light chain. *Proc. Natl Acad. Sci. USA* **88**, 10721–10725 (1991).
14. Wei, Z. & Lieber, M. R. Lymphoid V(D)J recombination. Functional analysis of the spacer sequence within the recombination signal. *J. Biol. Chem.* **268**, 3180–3183 (1993).
15. Connor, A. M. *et al.* Mouse V β 1783 recombination signal sequences mediate recombination more frequently than those of V β 1558. *J. Immunol.* **155**, 5268–5272 (1995).
16. Akira, S., Okazaki, K. & Sakano, H. Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science* **238**, 1134–1138 (1987).
17. Gauss, G. H. & Lieber, M. R. The basis for the mechanistic bias for deletional over inversive V(D)J recombination. *Genes Dev.* **6**, 1553–1561 (1992).
18. VanDyk, L. F., Wise, T. W., Moore, B. B. & Meek, K. Immunoglobulin D β recombination signal sequence targeting: effect of D β coding and flanking regions and recombination partner. *J. Immunol.* **157**, 4005–4015 (1996).
19. Chen, J., Lansford, R., Stewart, V., Young, F. & Alt, F. W. RAG-2-deficient blastocyst complementation: an assay of gene function in lymphocyte development. *Proc. Natl Acad. Sci. USA* **90**, 4528–4532 (1993).
20. Willerford, D. M., Swat, W. & Alt, F. W. Developmental regulation of V(D)J recombination and lymphocyte differentiation. *Curr. Opin. Genet. Dev.* **6**, 603–609 (1996).
21. Sikes, M. L., Suarez, C. C. & Oltz, E. M. Regulation of V(D)J recombination by transcriptional promoters. *Mol. Cell. Biol.* **19**, 2773–2781 (1999).

22. Doty, R. T., Xia, D., Nguyen, S. P., Hathaway, T. R. & Willerford, D. M. Promoter element for transcription of unrearranged T-cell receptor beta-chain gene in pro-T cells. *Blood* **93**, 3017–3025 (1999).
23. Whitehurst, C. E., Chattopadhyay, S. & Chen, J. Control of V(D)J recombinational accessibility of the DJ β gene segment at the TCR β locus by a germline promoter. *Immunity* **10**, 313–322 (1999).
24. Ramsden, D. A., Baetz, K. & Wu, G. E. Conservation of sequence in recombination signal sequence spacers. *Nucleic Acids Res.* **22**, 1785–1796 (1994).
25. Yancopoulos, G. D. *et al.* Preferential utilization of the most J μ -proximal V μ gene segments in pre-B-cell lines. *Nature* **311**, 727–733 (1984).
26. Jouvin-Marche, E. *et al.* Genomic organization of the T cell receptor V α family. *EMBO J.* **9**, 2141–2150 (1990).
27. Alt, F. W. *et al.* Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J.* **3**, 1209–1219 (1984).
28. Rock, E. P., Sibbald, P. R., Davis, M. M. & Chien, Y.-H. CRD3 length in antigen-specific immune receptors. *J. Exp. Med.* **179**, 323–328 (1994).
29. Sleckman, B. P., Bardou, C. G., Ferrini, R., Davidson, L. & Alt, F. W. Function of the TCR alpha enhancer in $\alpha\beta$ and $\gamma\delta$ T cells. *Immunity* **7**, 505–515 (1997).
30. Gärtner, F. *et al.* Immature thymocytes employ distinct signaling pathways for allelic exclusion versus differentiation and expansion. *Immunity* **10**, 537–546 (1999).
31. Sleckman, B. P. *et al.* Mechanisms that direct ordered assembly of T-cell receptor β locus V, D, and J gene segments. *Proc. Natl Acad. Sci. USA*. (in the press).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank O. Kanagawa and K. Murphy for advice and discussions and D. Fenoglio for assistance on cell sorting. This work is supported in part by the National Institutes of Health (F.W.A.). B.P.S. is a recipient of a Career Development Award from the Burroughs Wellcome Fund. C.H.B. was a fellow of the Irvington Institute for Immunological Research. F.G. is an associate of the Howard Hughes Medical Institute. F.W.A. is an investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to B.P.S. (e-mail: sleckman@immunology.WUSTL.edu).

Intraprotein radical transfer during photoactivation of DNA photolyase

Corinne Aubert*, Marten H. Vos†, Paul Mathis*, André P. M. Eker‡ & Klaus Brettel*

* Section de Bioénergétique (CNRS URA 2096), CEA Saclay, 91191 Gif-sur-Yvette Cedex, France

† INSERM U451, Laboratoire d'Optique Appliquée, Ecole Polytechnique-ENSTA, 91761 Palaiseau Cedex, France

‡ Department of Cell Biology and Genetics, Medical Genetics Centre, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands

Amino-acid radicals play key roles in many enzymatic reactions¹. Catalysis often involves transfer of a radical character within the protein, as in class I ribonucleotide reductase where radical transfer occurs over 35 Å, from a tyrosyl radical to a cysteine^{1–3}. It is currently debated whether this kind of long-range transfer occurs by electron transfer, followed by proton release to create a neutral radical, or by H-atom transfer, that is, simultaneous transfer of electrons and protons^{4–7}. The latter mechanism avoids the energetic cost of charge formation in the low dielectric protein^{4,5}, but it is less robust to structural changes than is electron transfer⁷. Available experimental data do not clearly discriminate between these proposals. We have studied the mechanism of photoactivation (light-induced reduction of the flavin adenine dinucleotide cofactor) of *Escherichia coli* DNA photolyase^{8–10} using time-resolved absorption spectroscopy. Here we show that the excited flavin adenine dinucleotide radical abstracts an electron from a nearby tryptophan in 30 ps. After subsequent electron transfer along a chain of three tryptophans,

the most remote tryptophan (as a cation radical) releases a proton to the solvent in about 300 ns, showing that electron transfer occurs before proton dissociation. A similar process may take place in photolyase-like blue-light receptors.

DNA photolyase catalyses the repair of major ultraviolet-induced DNA lesions, cyclobutane pyrimidine dimers or (6-4)-photoproducts, using the energy of near-ultraviolet light^{8,10} (Fig. 1). This enzyme, which is found in a variety of organisms ranging from bacteria to multicellular eukaryotes, consists of a single polypeptide with a relative molecular mass of ~55,000. It binds a flavin adenine dinucleotide (FAD) molecule as the essential catalytic cofactor, and it only functions when the FAD is fully reduced (FADH⁻). In isolated DNA photolyases, the FAD is typically in the inactive semireduced radical state FADH[•], but it can be easily reduced to the catalytically active FADH⁻ state by illumination with visible light in the presence of an exogenous electron donor, a process called photoactivation (Fig. 1).

The mechanism of photoactivation has been studied for *E. coli* photolyase^{9,11,12}. Excited FADH[•] oxidizes the solvent-accessible tryptophan W306, and exogenous electron donors then reduce the oxidized tryptophan (in competition with charge recombination occurring in about 10 ms; see Fig. 1, broken arrow) and thus stabilize the FADH⁻ state as required for the DNA repair function of the enzyme. We used absorption spectroscopy in the picosecond to millisecond time domain after selective excitation of FADH[•] by red laser flashes to study in detail the mechanism of radical transfer from FADH[•] to tryptophan (TrpH; H denotes the N1 proton) in recombinant *E. coli* photolyase¹³.

The final protonation state of the tryptophanyl radical was determined from absorption changes accompanying the recombination between FADH⁻ and the tryptophanyl radical in the absence of an exogenous electron donor (Fig. 2). Throughout the 400–645-nm spectral range studied, we observed absorbance changes appearing within 5 μ s and relaxing exponentially because of charge recombination with a time constant $\tau = 17$ ms (at pH 7.4). From the spectrum of these absorbance changes (Fig. 2a, spectrum A), we conclude that the tryptophanyl radical is in the unprotonated (neutral) state (Trp[•]) at times later than 5 μ s after excitation of FADH[•], because the difference between spectrum A and the (FADH⁻–FADH[•]) difference spectrum B is inconsistent with the absorption spectrum of the protonated (cation) radical TrpH^{•+}, but agrees well with that of Trp[•] (Fig. 2b). The recombination kinetics accelerated strongly upon lowering the pH ($\tau = 7$ ms at pH 6.5 and 0.9 ms at pH 5.4; see Fig. 2a, inset). The most straightforward explanation for the acceleration is that recombination requires Trp[•] reprotonation. The presence of Trp[•] is also consistent with the acidity of TrpH^{•+}, which has a $pK \approx 4$ in water^{14,15}. An electron

transfer from a tyrosine residue to Trp[•] as observed recently in photolyase from *Anacystis nidulans*^{16,17} was not detectable in photolyase from *E. coli*.

To establish whether Trp[•] is formed by hydrogen atom transfer or by electron transfer and subsequent deprotonation, we measured the kinetics of flash-induced absorbance changes in the nanosecond timescale in the 300–645-nm range where deprotonation of TrpH^{•+} is expected to give characteristic absorbance changes (see Fig. 2b). A typical signal measured at 580 nm is shown in the inset of Fig. 3a (trace H₂O). A fast, instrument-limited bleaching is followed by a further bleaching with $\tau \approx 300$ ns. The same kinetic components were observed at all other wavelengths studied. The spectrum of the amplitude of the 300-ns phase (Fig. 3a, spectrum A–B) agrees well

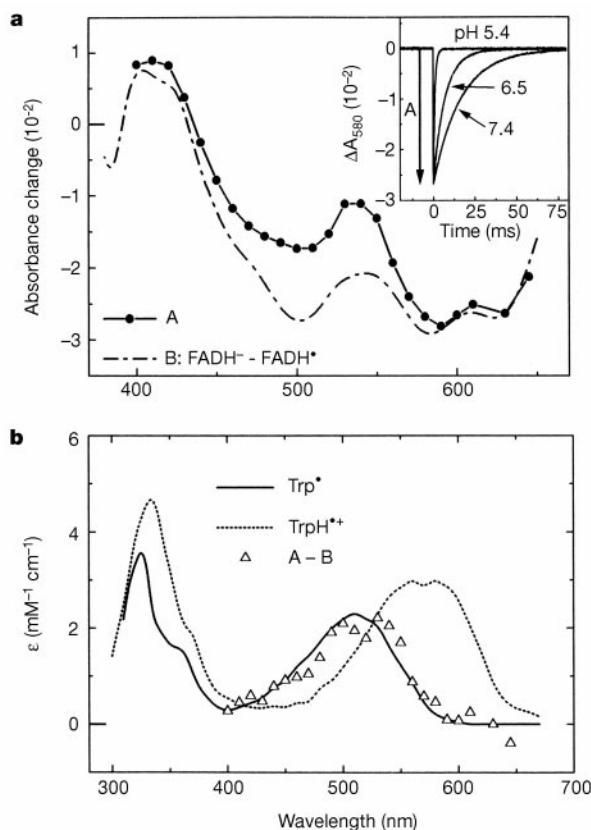


Figure 2 Evidence for a deprotonated tryptophanyl radical (Trp[•]). **a**, Trace A, spectrum of flash-induced absorbance changes decaying with $\tau = 17$ ms (pH 7.4), obtained from kinetic traces like that at pH 7.4 and 580 nm in the inset. Trace B, (FADH⁻–FADH[•]) difference spectrum in *E. coli* photolyase, obtained by subtracting absorption spectra recorded using a Kontron (model UV 922) spectrophotometer before and after saturating illumination with continuous light (540–650 nm) in the presence of 5 mM 1,4-dithiothreitol as exogenous electron donor. Spectrum B was normalized to spectrum A at ~630 nm. Inset, pH effect on the kinetics of the flash-induced absorbance change at 580 nm. The samples contained about 30 μ M photolyase, 200 mM NaCl, 15% (w/w) glycerol and 20 mM buffer (pH 7.4, Tris-HCl; pH 5.4 and pH 6.5, MES). Dithiothreitol was only added in spectrum B. Absorbance changes were recorded with a resolution of 5 μ s using continuous monitoring light¹⁶. The samples were excited by 7-ns dye laser pulses of ~30 mJ cm⁻² at 635 nm (for monitoring wavelengths $\lambda_m \leq 610$ nm) or ~20 mJ cm⁻² at 675 nm (for $\lambda_m \geq 610$ nm). Amplitudes in A were corrected for slight degradation of the samples during data collection and for the lower efficiency of the 675-nm excitation, using control measurements at 580 nm and 610 nm. **b**, Comparison of the difference between A and B (triangles) with absorption spectra of Trp[•] and TrpH^{•+} in aqueous solution (taken from ref. 15). The (A–B) difference spectrum was normalized to the other spectra at their isosbestic point (527 nm). The samples were kept at 10 °C during all spectroscopic measurements.

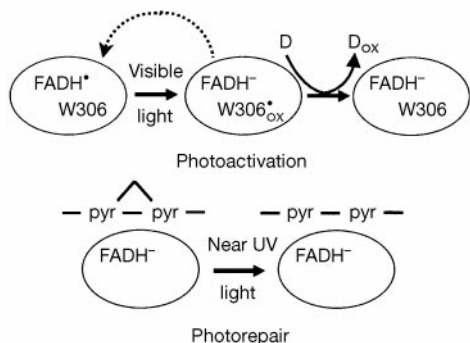


Figure 1 Outline of the photoactivation (top panel) and photorepair (bottom panel) reactions of DNA photolyase. See text for details. D, exogenous electron donor; pyr, pyrimidine in DNA.

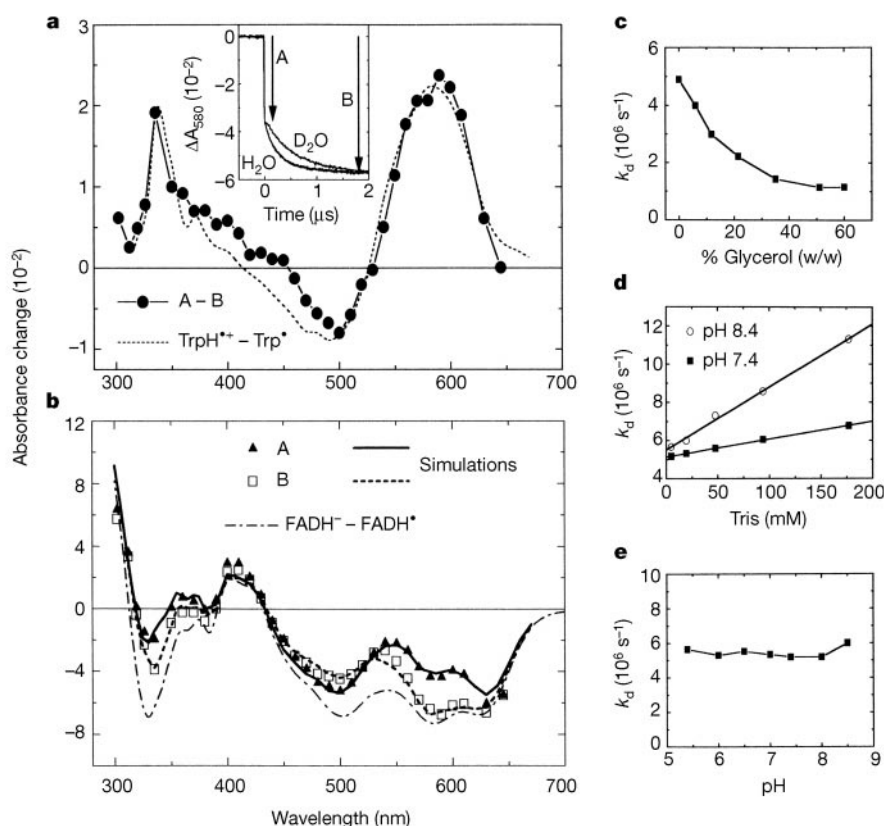


Figure 3 Monitoring the TrpH⁺⁺ deprotonation in the nanosecond time range. **a**, Circles, spectrum of the amplitude of the $\tau \approx 300$ -ns phase obtained from flash-induced absorbance changes like the 'H₂O' trace at 580 nm in the inset. Broken line, (TrpH⁺⁺ - Trp^{*}) difference spectrum calculated from the literature spectra in Fig. 2b and normalized to the spectrum of the 300-ns phase at ~ 580 nm. Inset, Deuterium isotope effect on the kinetics of the absorbance change at 580 nm. The H₂O sample contained 70 μ M photolyase, 200 mM NaCl, 10% (w/w) glycerol and 20 mM Tris, pH 7.4. The D₂O sample was obtained by exchange of H₂O for D₂O (ref. 17) and addition of 10% (v/v) glycerol. Absorbance changes were recorded with a time resolution of 10 ns as described¹⁶. Excitation conditions and correction procedure were as in Fig. 2. **b**, Triangles, spectrum of the absorbance changes preceding the 300-ns phase (amplitude A in the inset of **a**). This spectrum was simulated (solid line) by addition of the TrpH⁺⁺ spectrum (Fig. 2b) and the

(FADH⁻ - FADH^{*}) difference spectrum (dash-dotted line; for details, see Fig. 2a). Squares, spectrum of the absorbance changes after completion of the 300-ns phase, simulated (dashed line) by addition of the Trp^{*} spectrum (Fig. 2b) and the (FADH⁻ - FADH^{*}) difference spectrum. For the best simulations shown, the (FADH⁻ - FADH^{*}) differential molar absorption coefficient at 580 nm was assumed to be -2.4 times the molar absorption coefficient of TrpH⁺⁺ at 560 nm and -3.1 times that of Trp^{*} at 510 nm. **c-e**, Characterization of the deprotonation rate constant k_d of TrpH⁺⁺, obtained by least squares fits to flash-induced absorbance changes at 580 nm. Sample conditions: 30–70 μ M photolyase and 200 mM NaCl; and 20 mM Tris pH 7.4 and glycerol at the concentrations indicated (**c**); Tris pH 7.4 or pH 8.4 at the concentrations indicated (**d**); or 20 mM buffer (pH 5.4, 6 and 6.5, MES; pH 7, 7.5, 8 and 8.5, Tris) (**e**).

with the difference between the solution spectra of TrpH⁺⁺ and Trp^{*}, clearly indicating that the 300-ns phase reflects the deprotonation of TrpH⁺⁺. This conclusion is also supported by the kinetic isotope effect¹⁸ observed upon exchange of H₂O for D₂O: $\tau(\text{D}_2\text{O})/\tau(\text{H}_2\text{O}) = 2.2$ (see Fig. 3a, inset). As a consequence, the 300-ns phase should be preceded by electron transfer from TrpH to excited FADH^{*}, forming TrpH⁺⁺ and FADH⁻. Consistent with this prediction, the spectrum of the absorbance changes preceding the 300-ns phase (Fig. 3b, triangles) can be well simulated (solid line) by the addition of the TrpH⁺⁺ spectrum and the (FADH⁻ - FADH^{*}) difference spectrum.

We conclude that the radical transfer from excited FADH^{*} to the tryptophan residue W306 occurs by fast (< 10 ns) electron transfer (FADH^{*}..TrpH → FADH⁻..TrpH⁺⁺), followed by deprotonation of TrpH⁺⁺ in about 300 ns. The nature of each of these two reactions will now be examined in more detail.

With respect to the deprotonation of the tryptophanyl radical, it is particularly important to identify the proton acceptor. Increasing glycerol concentration in the solvent causes a pronounced decrease in the rate of TrpH⁺⁺ deprotonation (Fig. 3c), suggesting that the proton from TrpH⁺⁺ is released to the solvent rather than being

transferred to a proton accepting residue in the protein. The solvent contained two potential proton acceptors in sufficient concentration to account for a sub-microsecond reaction: unprotonated buffer molecules and H₂O itself. Increasing concentrations of Tris buffer accelerate the deprotonation of TrpH⁺⁺, and this acceleration is much more pronounced at pH 8.4 than at pH 7.4 (Fig. 3d), showing that the unprotonated Tris is the dominating proton acceptor at high concentrations of Tris. Extrapolation to zero Tris concentration yields a deprotonation rate constant, k_d , of about $5.3 \times 10^6 \text{ s}^{-1}$ which we attribute to proton transfer from TrpH⁺⁺ to H₂O in the solvent. Assuming $\text{pK} \approx 4$ for the deprotonating TrpH⁺⁺ in photolyase, this rate constant is consistent with the relationship $k_d = k_r \times 10^{-\text{pK}}$ for the deprotonation of weak acids in water, where k_r is the (diffusion controlled) reprotonation rate constant of typically 10^{10} – $10^{11} \text{ M}^{-1} \text{ s}^{-1}$ (see ref. 19 for a treatment of proton transfer processes in aqueous media). At the buffer concentration of 20 mM used in all our other experiments, H₂O should still be the dominating proton acceptor (see Fig. 3d). Varying the pH between 5.4 and 8.5 at 20 mM buffer, the kinetics of deprotonation (Fig. 3e) and its amplitude (data not shown) were essentially constant, as expected for H₂O ($\text{pK} = -1.7$) as proton acceptor.

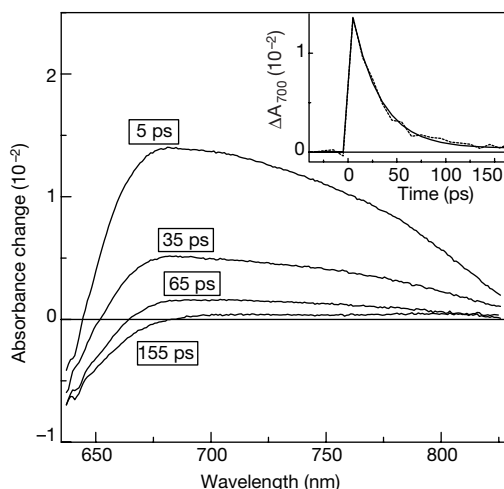


Figure 4 Study of the reaction $\text{FADH}^{\bullet\bullet} \cdot \text{TrpH} \rightarrow \text{FADH}^{\bullet} \cdot \text{TrpH}^{\bullet+}$ by ultrafast spectroscopy. Measurements were performed as described²⁶ with 620-nm pump pulses (70 fs full width at half maximum) and white-light continuum probe pulses, using a continuously moving thermostatted cell with an optical pathlength of 1 mm. Shown are transient spectra at selected delay times. Inset, kinetic data (dashed line) and exponential fit (solid line, $\tau = 29$ ps) at 700 nm. The sample contained 260 μM photolyase, 200 mM NaCl, 15% (w/w) glycerol and 20 mM Tris, pH 7.4.

The kinetics of the electron transfer reaction starting from the excited state $\text{FADH}^{\bullet\bullet}$ were resolved by ultrafast pump-probe spectroscopy in the near-infrared spectral region where an induced absorption has been assigned to $\text{FADH}^{\bullet\bullet}$ (refs. 11, 20). Figure 4 shows typical difference spectra recorded at given delay times, and the kinetic trace at 700 nm. The broad absorbance increase between 650 nm and 800 nm observed immediately after excitation (5-ps spectrum), attributed to $\text{FADH}^{\bullet\bullet}$, decays nearly completely with $\tau \approx 30$ ps (see 700-nm trace). The difference spectrum observed after completion of the 30-ps phase (see 155-ps spectrum) consists of a bleaching below 680 nm (the ground state absorption of

$\text{FADH}^{\bullet}$) and is consistent with the presence of the state $(\text{FADH}^{\bullet} \cdot \text{TrpH}^{\bullet+})$ (see Fig. 3b). These results show that the excited state $\text{FADH}^{\bullet\bullet}$ decays in about 30 ps, presumably by abstracting an electron from a nearby tryptophan residue (see below).

Figure 5 depicts a reaction scheme for the radical transfer from excited $\text{FADH}^{\bullet}$ to W306 that is based on our results and takes into account essential elements of the X-ray structure of *E. coli* photolyase²¹. The solvent accessible tryptophan residue W306 is as far as 13.4 Å (shortest edge-to-edge distance) from the flavin, excluding a direct or superexchange-mediated electron transfer to $\text{FADH}^{\bullet\bullet}$ in 30 ps (refs. 7, 22). However, two other tryptophan residues, W382 and W359, are located in between (Fig. 5, inset), suggesting an electron transfer in three steps. The first step, downhill electron transfer from W382 to $\text{FADH}^{\bullet\bullet}$ takes 30 ps, the observed lifetime of $\text{FADH}^{\bullet\bullet}$ (Fig. 4). The next two steps, the presumably nearly isoenergetic electron transfers from W359 to the cation radical of W382 and from W306 to the cation radical of W359, should be faster than 10 ns because otherwise the deprotonation of the cation radical of W306, measured with a time resolution of 10 ns (Fig. 3a), should have shown a lag phase. Given the short distances between the partners (Fig. 5, inset), these three electron transfer rates are consistent with established rates of other intraprotein electron transfer reactions⁷. The proton release from the cation radical of W306 to the aqueous medium in about 300 ns is driven by a substantial decrease in free energy (about 0.2 eV at pH 7.4, assuming a pK of 4 for $\text{TrpH}^{\bullet+}$). This energetic stabilization is essential to trap the oxidizing radical on W306, which is accessible to exogenous reductants and sufficiently far apart from $\text{FADH}^{\bullet}$ to prevent a fast direct recombination; moreover, the deprotonation creates a free energy barrier in a potential recombination pathway through repopulation of the cation radical of W306 and reverse electron transfer in the chain $\text{FAD} - \text{W382} - \text{W359} - \text{W306}$ (Fig. 5, broken arrows). The acceleration of the recombination with decreasing pH (Fig. 2a, inset) indicates that this is the actual recombination pathway.

Examination of the X-ray structure of *E. coli* photolyase²¹ reveals important differences between the environments of the three tryptophans. The indole ring of W306 is partially surrounded by mainly polar amino acids. Its ring nitrogen is directly exposed to a shallow pocket that is open to the solvent and that can accommodate

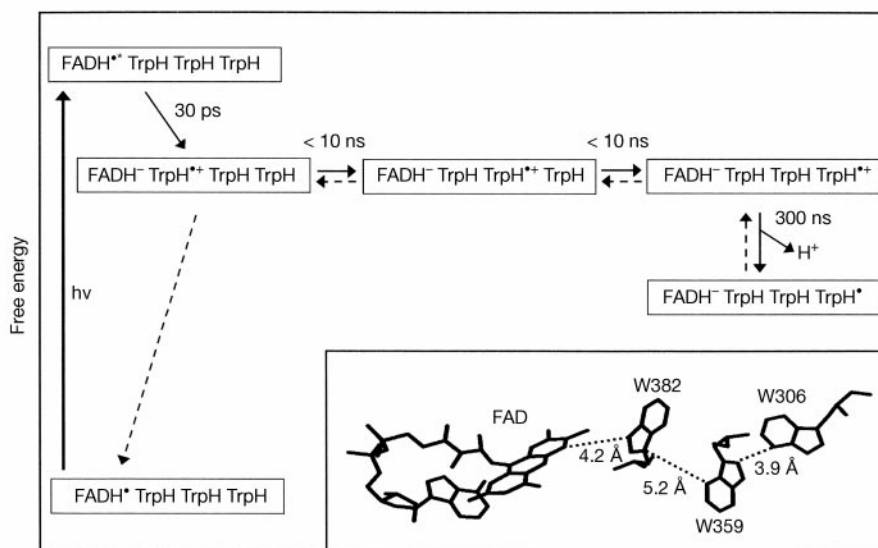


Figure 5 Reaction scheme for the radical transfer during photoactivation of *E. coli* photolyase. TrpH TrpH TrpH refers (from left to right) to W382, W359 and W306. Productive (forward) reactions are indicated as solid arrows, and back reactions as broken arrows. The free energy levels are not drawn to scale. We estimate the excited state of $\text{FADH}^{\bullet}$ to be ~ 2.0 eV above the ground state (corresponding to the energy of a 630-nm

photon), and the $(\text{FADH}^{\bullet} \cdot \text{TrpH}^{\bullet+})$ states to be ~ 1.5 eV (assuming midpoint potentials of about -0.4 V for $\text{FADH}^{\bullet}/\text{FADH}^-$ in photolyase²⁷ and 1.1 V for $\text{TrpH}^{\bullet+}/\text{TrpH}$ buried in a protein¹⁴). Inset, arrangement and shortest edge-to-edge distances between FAD, W382, W359 and W306 obtained from the X-ray structure of photolyase from *E. coli*²¹. Swiss-PdbViewer²⁸ was used to prepare the inset.

a water molecule, a situation that should favour proton transfer from the W306 cation radical to the solvent. In contrast, W382 is buried in the hydrophobic interior of the protein. There is no potential proton acceptor within 5 Å of the ring nitrogen of W382, excluding the possibility of deprotonation of the cation radical. The indole ring of W359 is surrounded by both hydrophobic and polar amino acids. The ring nitrogen appears to be hydrogen bonded to a more buried water molecule that might assist deprotonation of the W359 cation radical; however, our experimental data with a time resolution of 10 ns, together with previous data on the W306F mutant¹², do not give any indication for a deprotonation of the W359 cation radical.

The different polarities of the tryptophan environments should be relevant for the energetics of the electron transfer, as the reduction potential of the TrpH⁺/TrpH couple is higher in a hydrophobic protein environment than in water¹⁴. For this reason, there is likely to be some decrease of the free energy of the TrpH⁺ states in the order W382, W359, W306, favouring radical localization on W306 already, before its deprotonation.

A reaction scheme quite different from Fig. 5 was proposed previously⁹. It assumed that a long-lived quartet state of FADH[•] (formed by intersystem crossing from the initial excited doublet state) abstracts an electron from W306 in about 500 ns, and that the resulting cation radical TrpH^{•+} remains protonated. This mechanism is incompatible with our results which establish unambiguously that TrpH^{•+} is formed in less than 10 ns and deprotonates in about 300 ns. The origin of the disagreement is most probably that the previous scheme was based on studies^{11,20,23} using 355-nm laser flashes or Xe flashes that could excite species other than FADH[•] (for example, fully oxidized FAD; see ref. 24). Under our excitation conditions, FADH[•] is the only species being excited.

In summary, we have reached a consistent mechanism for the radical transfer during photoactivation of photolase in which sequential electron transfer among amino-acid residues generates transient cation radicals before charge neutralization by proton release to the solvent. Thus, we conclude that charge compensating simultaneous proton transfer is not a prerequisite for intraprotein radical transfer. A definitive answer to the question of whether long-range radical transfer in other enzymes proceeds through a similar mechanism may require time-resolved methods to selectively monitor electron transfer, H-atom transfer and protonation/deprotonation events.

Our results on the activation of DNA photolase from *E. coli* may be relevant for the other photolases (class I and class II)¹⁰ and also for the cryptochrome blue-light receptors in plants and animals, which are flavoproteins similar in sequence to photolases²⁵. Indeed, inspecting the aligned amino-acid sequences¹⁰ of known photolases and cryptochromes, we noticed that the tryptophan residues W382, W359 and W306 are uniformly conserved. This observation suggests that the electron transfer chain described here is involved in the function of all photolases, and of blue light receptors as well. □

Received 30 December 1999; accepted 10 January 2000.

1. Stubbe, J. & van der Donk, W. A. Protein radicals in enzyme catalysis. *Chem. Rev.* **98**, 705–762 (1998).
2. Uhlin, U. & Eklund, H. Structure of ribonucleotide reductase protein R1. *Nature* **370**, 533–539 (1994).
3. Sjöberg, B.-M. The ribonucleotide reductase jigsaw puzzle: a large piece falls into place. *Structure* **2**, 793–796 (1994).
4. Siegbahn, P. E. M., Blomberg, M. R. A. & Crabtree, R. H. Hydrogen transfer in the presence of amino acid radicals. *Theor. Chem. Acc.* **97**, 289–300 (1997).
5. Siegbahn, P. E. M., Eriksson, L., Himo, F. & Pavlov, M. Hydrogen atom transfer in ribonucleotide reductase (RNR). *J. Phys. Chem.* **102**, 10622–10629 (1998).
6. Cukier, R. I. & Nocera, D. G. Proton coupled electron transfer. *Annu. Rev. Phys. Chem.* **49**, 337–369 (1998).
7. Page, C. C., Moser, C. C., Chen, X. & Dutton, P. L. Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature* **402**, 47–52 (1999).
8. Sancar, A. No “end of history” for photolases. *Science* **272**, 48–49 (1996).
9. Kim, S. -T., Heelis, P. F. & Sancar, A. Role of tryptophans in substrate binding and catalysis by DNA photolase. *Methods Enzymol.* **258**, 319–343 (1995).

10. Yasui, A. & Eker, A. P. M. in *DNA Damage and Repair*, Vol. 2: *DNA Repair in Higher Eukaryotes* (eds Nickoloff, J. A. & Hoekstra, M. E.) 9–32 (Humana, Totowa, 1998).
11. Heelis, P. F., Okamura, T. & Sancar, A. Excited-state properties of *Escherichia coli* DNA photolase in the picosecond to millisecond time scale. *Biochemistry* **29**, 5694–5698 (1990).
12. Li, Y. F., Heelis, P. F. & Sancar, A. Active site of DNA photolase: tryptophan-306 is the intrinsic hydrogen atom donor essential for flavin radical photoreduction and DNA repair *in vitro*. *Biochemistry* **30**, 6322–6329 (1991).
13. Eker, A. P. M., Yajima, H. & Yasui, A. DNA photolase from the fungus *Neurospora crassa*. Purification, characterization and comparison with other photolases. *Photochem. Photobiol.* **60**, 125–133 (1994).
14. Tommos, C., Skaliky, J. J., Pilloud, D. L., Wand, A. J. & Dutton, P. L. *De novo* proteins as models of radical enzymes. *Biochemistry* **38**, 9495–9507 (1999).
15. Solar, S., Getoff, G., Surdhar, P. S., Armstrong, D. A. & Singh, A. Oxidation of tryptophan and N-methylindole by N₂⁺, Br₂^{•+}, and (SCN)₂^{•+} radicals in light- and heavy-water solutions: A pulse radiolysis study. *J. Phys. Chem.* **95**, 3636–3643 (1991).
16. Aubert, C., Mathis, P., Eker, A. P. M. & Brettel, K. Intraprotein electron transfer between tyrosine and tryptophan in DNA photolase from *Anacystis nidulans*. *Proc. Natl Acad. Sci. USA* **96**, 5423–5427 (1999).
17. Aubert, C., Brettel, K., Mathis, P., Eker, A. P. M. & Boussac, A. EPR detection of the transient tyrosyl radical in DNA photolase from *Anacystis nidulans*. *J. Am. Chem. Soc.* **121**, 8659–8660 (1999).
18. Cleland, W. W., O’Leary, M. H. & Northrop, D. B. *Isotope Effects on Enzyme-Catalyzed Reactions* (Univ. Park Press, Baltimore, London, Tokyo, 1977).
19. Eigen, M. Proton transfer, acid-base catalysis, and enzymatic hydrolysis Part I: elementary processes. *Angew. Chem. Internat. Edn.* **3**, 1–19 (1964).
20. Okamura, T., Sancar, A., Heelis, P. F., Hirata, Y. & Mataga, N. Doublet-quartet intersystem crossing of flavin radical in DNA photolase. *J. Am. Chem. Soc.* **111**, 5967–5969 (1989).
21. Park, H. W., Kim, S. T., Sancar, A. & Deisenhofer, J. Crystal structure of DNA photolase from *Escherichia coli*. *Science* **268**, 1866–1872 (1995).
22. Cheung, M. S., Daizadeh, I., Stuchebrukhov, A. A. & Heelis, P. F. Pathways of electron transfer in *Escherichia coli* DNA photolase: Trp³⁰⁶ to FADH. *Biophys. J.* **76**, 1241–1249 (1999).
23. Kim, S. T., Sancar, A., Essenmacher, C. & Babcock, G. T. Time-resolved EPR studies with DNA photolase: excited-state FADH[•] abstracts an electron from Trp-306 to generate FADH[•], the catalytically active form of the cofactor. *Proc. Natl Acad. Sci. USA* **90**, 8023–8027 (1993).
24. Gindt, Y. M. *et al.* Origin of the transient electron paramagnetic resonance signals in DNA photolase. *Biochemistry* **38**, 3857–3866 (1999).
25. Cashmore, A. R., Jarillo, J. A., Xu, Y. -J. & Liu, D. Cryptochromes: blue light receptors for plants and animals. *Science* **284**, 760–765 (1999).
26. Martin, J.-L. & Vos, M. H. Femtosecond measurements of geminate recombination in heme proteins. *Methods Enzymol.* **232**, 416–430 (1994).
27. Heelis, P. F., Deeble, D. J., Kim, S. T. & Sancar, A. Splitting of cys-syn cyclobutane thymine–thymine dimers by radiolysis and its relevance to enzymatic photoreactivation. *Int. J. Radiat. Biol.* **62**, 137–143 (1992).
28. Guex, N. & Peitsch, M. N. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).

Acknowledgements

We thank A. Yasui for providing us with the *E. coli* photolase expression construct; P. Barth and P. L. Dutton for helpful discussions; and T. A. Mattioli for critical reading of the manuscript. M.H.V. is supported by CNRS.

Correspondence and requests for materials should be addressed to K.B. (e-mail: brettel@dsvidf.cea.fr).

Engineering stability in gene networks by autoregulation

Attila Becskei & Luis Serrano

EMBL, Structures & Biocomputing, Meyerhofstrasse 1, Heidelberg D-69012, Germany

The genetic and biochemical networks which underlie such things as homeostasis in metabolism and the developmental programs of living cells, must withstand considerable variations and random perturbations of biochemical parameters^{1–3}. These occur as transient changes in, for example, transcription, translation, and RNA and protein degradation. The intensity and duration of these perturbations differ between cells in a population⁴. The unique state of cells, and thus the diversity in a population, is owing to the different environmental stimuli the individual cells experience and the inherent stochastic nature of biochemical processes (for example, refs 5 and 6). It has been proposed, but not

demonstrated, that autoregulatory, negative feedback loops in gene circuits provide stability⁷, thereby limiting the range over which the concentrations of network components fluctuate. Here we have designed and constructed simple gene circuits consisting of a regulator and transcriptional repressor modules in *Escherichia coli* and we show the gain of stability produced by negative feedback.

To test the role of negative feedback in the stability of gene networks, we first designed simple gene circuits based on simple control systems (Fig. 1c, d). To construct the autoregulatory system, the tetracycline repressor (TetR) of the transposon Tn10 was fused to the green fluorescent protein (EGFP) (TetR-EGFP) and placed downstream of the lambda promoter containing two tetracycline operators (Fig. 2a). As controls, the unregulated counterparts were obtained by slight modification of this system. In this way, differences among the systems could be neglected, while the feedback was eliminated (Fig. 2b, c).

To get a quantitative measure of the stability of our systems, a linear stability analysis was performed using differential equation models of gene circuits. The use of quantitative stability analysis allows modelling of biological processes in their natural context. When a particular process is selected for examination *in vivo*, its description is biased by the impact of the unique state of the cell, so it cannot be fully characterized as a separate entity isolated from the perturbations of the biochemical network of the cell. Actually, comparative stability analysis takes the cellular perturbations into account. The value of the stability (S) is obtained by the linearization of the equations around the steady state. The absolute values of stability of the unregulated (1) and autoregulatory (2) systems were compared as the ratio of the stability (3) of the two systems.

$$S_{\text{unreg}} = f'_{\text{unreg}}(R^*) = -k_{\text{deg}} \quad (1)$$

$$S_{\text{auto}} = f'_{\text{auto}}(R^*) = -\frac{nk_p P k_i a k_r}{(1 + k_p P + k_i R^*)^2} - k_{\text{deg}} \quad (2)$$

$$S_r = \frac{S_{\text{auto}}}{S_{\text{unreg}}} \quad (3)$$

Where R is the concentration of the repressor (R^* at the steady state), P is the concentration of the RNA polymerase, k_p and k_r are the binding constants of the polymerase and the repressor, respectively, k_i is the promoter isomerization rate from closed to initiating complex, a is the proportionality constant between the mRNA and protein concentration, k_{deg} is the degradation rate of the repressor and n is the gene copy number.

It is apparent that for all positive values of parameters and steady state concentrations the stability is higher in the autoregulatory system. For the biochemical parameters of the systems constructed here, the autoregulatory system shows a twofold increase in stability over the unregulated one. Varying the parameters within a range of values with biological significance demonstrates that the stability gain by negative feedback is preserved. However, it decreases considerably when the binding constant for the repressor or the transcription rate is very low, especially when the repressor has a short half-life (Fig. 1a, b). Actually, low repression factor and low concentration of repressor affect the essential property of negative feedback. A detailed mathematical analysis shows that conditions can be found when the degree of stability of the two systems converges while the parameters are unchanged. However, the autoregulatory system is generally superior, and only equal to the unregulated system under certain conditions. To simulate the unique perturbation states of cells, random perturbations were applied to the steady state in a cell population, which resulted in a variation of the actual concentration of the transcription factor around the steady state in the population (Fig. 1). However, the

distribution is narrower in the autoregulatory system owing to the higher stability.

In the second step, we analysed the degree of variability in the expression of the TetR-EGFP by fluorescence microscopy in the autoregulatory and three unregulated systems. Coefficient of variation (V_c) was used as a measure of variability. The expression of the TetR-EGFP in the autoregulatory loop is characterized by a low steady-state level and high degree of homogeneity (Fig. 2a), with a V_c of 6–9% (Fig. 3a, column A). The narrow distribution is not specific to very low copy plasmids. Essentially the same V_c values were observed in cell populations harbouring very low, middle and high copy number plasmids (data not shown).

The negative feedback is expected to be maintained if a low concentration of inducer is added to the medium. When anhydrotetracycline (atc) is added at very low concentration (1 ng ml^{-1}), V_c stays below 10%. However, a transition from the autoregulatory to the first unregulated system occurs at higher inducer concentrations owing to the reduction of feedback repression (Fig. 3b).

As predicted, decreasing the affinity of the repressor by a mutation⁸ increases the variability in the second system (Fig. 2b) and the steady-state concentration of the fusion protein. The variability was comparable to that encountered in the autoregulatory system when high concentrations of inducer were added ($\sim 20\%$, Fig. 3a). The significant difference of the steady-state concentrations is because of the high binding affinity of the wild-type TetR, which is also involved in the stabilization of feedback.

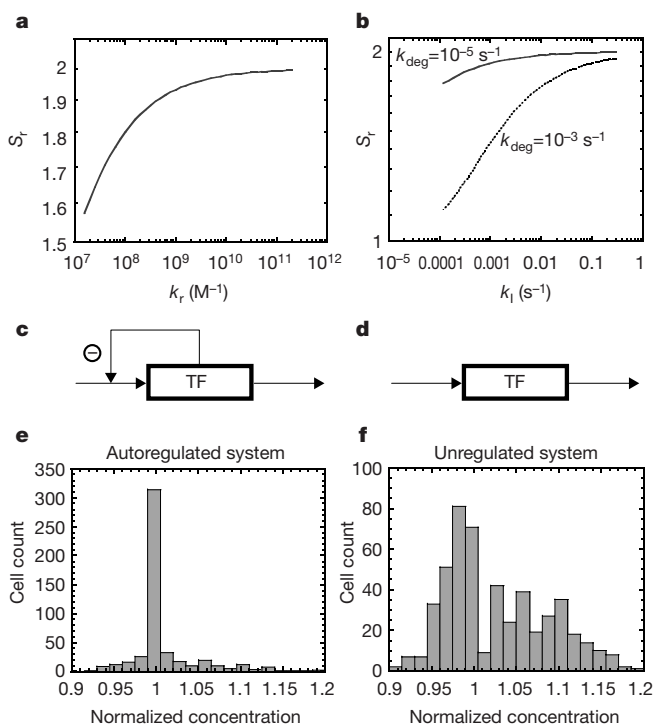


Figure 1 Stability properties of gene circuits. **a, b**, For calculating the value of relative stability (S_r , equation 3), the parameters of the system are: $P = 100 \text{ nM}$, $k_p = 1.5 \times 10^{10} \text{ M}^{-1}$, $k_i = 0.3 \text{ nM s}^{-1}$, $n = 3$, $a = 3.3$, $k_r = 2 \times 10^{11} \text{ M}^{-1}$, $k_{\text{deg}} = 10^{-5} \text{ s}^{-1}$, while one parameter is selected for independent variable: k_r (**a**) or k_{deg} (**b**). In *E. coli* k_i ranges from 10^{-4} to 1 synthesized RNAs²⁴. **c–d**, Control theoretical diagrams for negative feedback (**c**) and unregulated (**d**) systems containing a transcription factor (TF). Inflow arrow stands for the rate of synthesis, the outflow arrow for the rate of degradation. **e–f**, Numerical simulation is given for both systems, where the actual concentration of each cell is normalized by the calculated steady-state value.

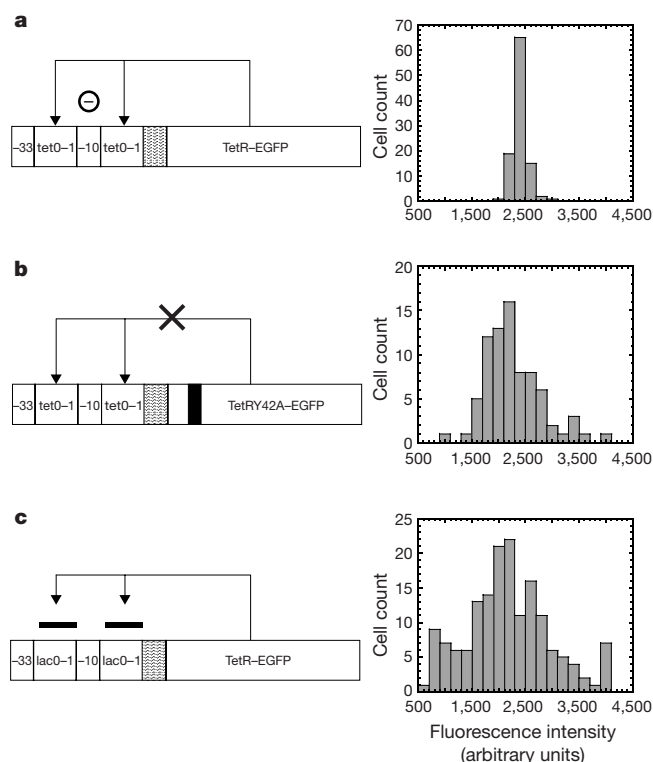


Figure 2 Gene circuits and corresponding typical distributions of fluorescence intensities. Each circuit consists of the following units: operator at position V, -33 hexamer, operator at position IV, -10 hexamer, untranslated region, fusion protein. GFP fluorescence intensity is denoted by arbitrary units in the histograms which corresponds to intensities of a 12-bit (4,096 grey level) image. The mean values of the three systems were normalized from different exposure times. Relative mean values (r.m.v.) are given for each distribution. **a**, Autoregulatory system in DH5 α cells; r.m.v. = 1. **b**, The repressor was mutated (Y42A) in the DNA-binding domain to obtain the unregulated system in DH5 α cells; r.m.v. = 38. **c**, Unregulated system obtained by operator replacement in DH5 α Z1 cells. Sample is taken after induction by IPTG; r.m.v. = 14.

In both systems, there is a residual binding of the repressor to the operator, as both the TetR-atc complex and the TetRY42A mutant have low binding affinities to the operator^{8,9}. As a third model for an unregulated system, the tet operator was replaced by the lac operator (Fig. 2c), so only a negligible nonspecific protein-DNA interaction remains. The expression of TetR-EGFP was induced by saturating isopropyl- β -D-thiogalactopyranoside (IPTG) concentration (1 mM) to minimize stochastic induction phenomena observed at subsaturating concentrations of inducer^{6,10}. The distribution of the fluorescence intensities was broader than in any other construct (Fig. 2c); two hours after induction V_c was above 30% (Fig. 3b). Furthermore, we were able to examine whether long-term expression affects the variability. High variability was observed in both the long-term expression of the TetRY42A-EGFP mutant, where the steady state is continuously maintained during bacterial growth, and in the short-term expression of TetR-EGFP, where expression is induced by release of Lac repressor from the operators (Fig. 3b).

Although V_c is appropriate to compare the variation of cell populations independent of the magnitude of their means, we checked that the homogeneity is not merely a consequence of the low steady-state level of TetR-EGFP in the autoregulatory system. This was done by two different experiments. First, samples were taken at different time intervals shortly after induction from the third unregulated system (Fig. 2c). Samples were selected that have

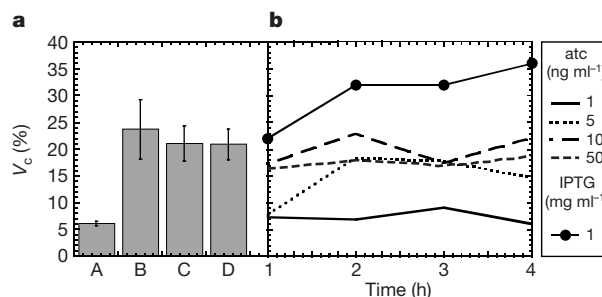


Figure 3 Variability in classical and autoregulatory systems. **a**, Mean coefficients of variation were calculated from three independent experiments using independent transformants, and were considered as simple statistical variables to calculate standard errors indicated on error bars. This characterizes the range of variability of independent experiments. Column A, the autoregulatory system; column B, an EGFP expression vector under the regulation of chromosomal TetR; column C, the operator-replaced unregulated system (Fig. 2c) after addition of 1 mM IPTG; column D, the mutant-repressor unregulated system (Fig. 2b). The same exposure times were applied to take the images for statistical analysis. **b**, Variability in the course of time after induction. The autoregulatory system is induced by different atc concentrations. The unregulated system (Fig. 2c) is induced by IPTG. The scale for V_c is the same in **a** and **b**.

similar mean fluorescence intensities to those in the autoregulatory system. There was a considerable difference in variability between the unregulated ($V_c \approx 20\%$) and the autoregulatory systems ($V_c \approx 6\%$) while their means were equal (Fig. 3a). Second, we measured EGFP expression under the control of chromosomal TetR. TetR controlled vector systems provide the most tightly regulated and evenly distributed expression presently available¹¹. At low inducer concentrations (3–5 ng ml⁻¹ atc), the mean EGFP fluorescence intensity of the cell population is roughly equal to that of the autoregulatory system. However, it shows about three-fold higher variability compared with the autoregulatory system (Fig. 3a).

Finally we investigated the possibility of dosage compensation by negative feedback loop. We monitored the mean expression level of luciferase produced by TetR-EGFP/luc bicistronic transcription unit when using low, medium and high copy plasmids (3–4, 20–30, 50–70 copies per cell, respectively). When the mean luminescence activity of the low copy system is normalized to one, the values for middle and high copy system are 2.6 ± 0.26 and 4.8 ± 0.69 , respectively. These data could be explained by the analytical description of the steady state, which is a power function of the plasmid copy number. Thus, autoregulation might contribute to dosage compensation but additional mechanisms are necessary to keep the protein at a constant concentration, independent of gene copy number. Indeed, both negative feedback and postranscriptional regulation participate in the dosage compensation of histone genes in yeast^{12,13}.

Recent works of integrative theoretical-experimental approach have shed light on some of the basic principles of the functioning of gene networks^{14,15}. Here we have focused on the stability by autoregulation proving its superiority to unregulated systems. In fact, roughly 40% of known transcription factors in *E. coli* negatively autoregulate themselves¹⁶. Reduced or increased amounts of the same transcriptional regulator can be the cause of severe developmental disorders¹⁷. Negative feedback provides a mechanism to ensure a homogeneous distribution of a transcriptional repressor within optimal concentration limits. Under certain conditions, however, the stability of the autoregulatory system could approach that of the classical system. Variability can be advantageous for adaptation¹⁸, stability for homeostasis. □

Methods

Analysis of stability

Equations 4 and 5 are based on thermodynamic-kinetic principles¹⁹ and the following data:

$$f_{\text{unreg}}(R) = \frac{dR}{dt} = n \frac{k_p P}{1 + k_p P} k_i a - k_{\text{deg}} R \quad (4)$$

$$f_{\text{auto}}(R) = \frac{dR}{dt} = n \frac{k_p P}{1 + k_p P + k_i R} k_i a - k_{\text{deg}} R \quad (5)$$

(1) The TetR exists only in the form of dimers in bacterial cells owing to the very high dimerization constant ($k_d = 4.8 \times 10^{13}$; ref. 20), so a linear correlation can be assumed between the concentration of TetR transcript and that of the dimer. Thus, the dimer concentration is half of that of the transcript. (2) Only the operator at position IV contributes significantly to repression. Its repression factor is around 100 times higher than that of the operator at position V (see legend of Fig. 2). The strength of the promoters containing the tetO-1 and lacO-1 operators are approximately the same²¹. (3) The RNA polymerase and the TetR bind mutually exclusively to the operator (V)²¹. The steady-state level of the repressor (R^* when $f(R) = 0$) is a linear function of gene copy number in unregulated systems, but a square-root function in autoregulatory systems. The stability value is obtained by the Taylor expansion of the $f(R^* + \eta)$ where η is the perturbation: $f(R^* + \eta) = f(R^*) + \eta f'(R^*) + O(\eta^2)$. $O(\eta^2)$ denotes quadratically small terms in η and can be neglected. The magnitude of $f'(R^*)$ determines the decay rate of the perturbation²². For numerical simulations, random values of same relative range were taken for η to perturb the steady state of the circuits at random time intervals, and the actual concentration of the transcription factor was taken from a population of 500 cells.

Strains, growth conditions and media

The *E. coli* DH5 α and DH5 α Z1 strains were used in all experiments. DH5 α Z1 is derived from DH5 α by chromosomal integration of transcriptional repressors LacI and TetR and also a spectinomycin resistance (Sp^r) marker²³. DH5 α Z1 produces $\sim 3,000$ molecules of LacI and $\sim 7,000$ molecules of TetR per cell²³. Cells were grown in Luria broth with $30 \mu\text{g ml}^{-1}$ kanamycin and $50 \mu\text{g ml}^{-1}$ streptomycin. Cells from overnight cultures were diluted 1:100 and grown at 25°C . Expression was induced at early log phase by addition of different atc ($1\text{--}100 \text{ ng ml}^{-1}$) concentrations or IPTG at saturating concentration (1 mM) depending on the system. Samples for examination were taken from mid-log cultures.

Gene circuit construction

TetR, EGFP and firefly luciferase were amplified by polymerase chain reaction. The TetRY42A mutant was created by site-directed mutagenesis introducing a GCC triplet coding for alanine and verified by sequencing. The fusion protein is constructed from amino terminally positioned TetR or TetRY42A, a linker (with amino-acid sequence GENLYFQSGGA) and a carboxy terminally positioned EGFP. The tetO-1 and the lacO-1 regulatory units were cloned from pZE21MCS-1 and pZE22MCS-1. The regulatory and repressor modules were cloned into the very low copy plasmid pZS*21MCS-1 and in the case of the autoregulatory system the origin of replication was also changed to p15A and ColE1 to yield middle and high copy numbers per cell. To construct the bicistronic transcription unit the upstream TetREGFP module was ligated to luciferase separated by a ribosomal binding site and luciferase activity was measured as described²³.

Microscopy and statistical analysis

Bacteria were pelleted, washed and resuspended in PBS containing $50 \mu\text{g ml}^{-1}$ chloramphenicol to stop further protein synthesis. For microscopic examination, $10 \mu\text{l}$ of 1.5% agar was dropped on the microscope slide to form a thin layer. Immediately after drying, $1 \mu\text{l}$ of cell suspension was pipetted onto it and covered with a coverslip. Microscopy was carried out using a Leica DMIRB/E microscope equipped with a highly light-sensitive Hamamatsu C4742-95 digital CCD camera and an automatic light shutter. A Chroma GFP 513852 photocube transmitting a wavelength of $485\text{--}495 \text{ nm}$ was used to stimulate GFP fluorescence. Fluorescent images of single fields of cells were obtained using a phase contrast $\times 63/1.32$ oil-immersion lens with exposure times in a range between 10 and 4,000 ms. No detectable autofluorescence of control cells was observed in this range. Multiple exposures were taken of each field to assure that the fluorescence signal was within the range of the CCD camera. To compare variability of different systems having approximately the same mean fluorescence intensity, images were taken with the same exposure time on the same day. Relative mean values (r.m.v.) were calculated as the ratios of exposure times of two measurements. Transmission images were taken to identify all bacterial cells selected for fluorescence intensity measurement. For statistical analysis,

mean GFP fluorescence intensities were measured using the Openlab 2.0.6 software. The mean fluorescence intensities of 75 to 150 individual cells from the central areas of single fields were quantified. To compare standard deviations in broad range of fluorescence intensities, we calculated coefficients of variation (V_c), which is not biased by the photobleaching of GFP. V_c is the simple ratio of standard deviation to mean at high sample numbers. Coefficients of variation of the same field were compared for different exposure times within the linear range. They are within the range of the standard error of V_c and are independent of exposure time and background fluorescence. The standard error of V_c is small for large samples.

Received 25 December 1999; accepted 13 April 2000.

- Little, J. W., Shepley, D. P. & Wert, D. W. Robustness of a gene regulatory circuit. *EMBO J.* **18**, 4299–4307 (1999).
- Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
- Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
- Ko, M. S., Nakauchi, H. & Takahashi, N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *EMBO J.* **9**, 2835–2842 (1990).
- Nutt, S. L. *et al.* Independent regulation of the two Pax5 alleles during B-cell development. *Nat. Genet.* **21**, 390–395 (1999).
- Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc. Natl Acad. Sci. USA* **43**, 553–566 (1957).
- Savageau, M. A. Comparison of classical and autogenous systems of regulation in inducible operons. *Nature* **252**, 546–549 (1974).
- Baumeister, R., Helbl, V. & Hillen, W. Contacts between Tet repressor and tet operator revealed by new recognition specificities of single amino acid replacement mutants. *J. Mol. Biol.* **226**, 1257–1270 (1992).
- Lederer, T. *et al.* Tetracycline analogs affecting binding to Tn10-Encoded Tet repressor trigger the same mechanism of induction. *Biochemistry* **35**, 7439–7446 (1996).
- Siegle, D. A. & Hu, J. C. Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc. Natl Acad. Sci. USA* **94**, 8168–8172 (1997).
- Blau, H. M. & Rossi, F. M. Tet B or not tet B: advances in tetracycline-inducible gene expression. *Proc. Natl Acad. Sci. USA* **96**, 797–799 (1999).
- Moran, L., Norris, D. & Osley, M. A. A yeast H2A-H2B promoter can be regulated by changes in histone gene copy number. *Genes Dev.* **4**, 752–763 (1990).
- Osley, M. A. & Hereford, L. M. Yeast histone genes show dosage compensation. *Cell* **24**, 377–384 (1981).
- Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
- Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
- Thieffry, D., Huerta, A. M., Perez-Rueda, E. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* **20**, 433–440 (1998).
- Schedl, A. *et al.* Influence of PAX6 gene dosage on development: overexpression causes severe eye abnormalities. *Cell* **86**, 71–82 (1996).
- Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
- Wolf, D. M. & Eeckman, F. H. On the relationship between genomic regulatory element organization and gene regulatory dynamics. *J. Theor. Biol.* **195**, 167–186 (1998).
- Backes, H. *et al.* Combinations of the alpha-helix-turn-alpha-helix motif of TetR with respective residues from LacI or 434Cro: DNA recognition, inducer binding, and urea-dependent denaturation. *Biochemistry* **36**, 5311–5322 (1997).
- Lanzer, M. & Bujard, H. Promoters largely determine the efficiency of repressor action. *Proc. Natl Acad. Sci. USA* **85**, 8973–8977 (1988).
- Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications in Physics, Biology, Chemistry, and Engineering* (Perseus, Boulder, CO, 1994).
- Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1–12 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
- Record JR., M. T., Reznikoff, W. S., Craig, M. L., McQuade, K. L. & Schlax, P. J. *E. coli* and *S. typhimurium: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 792–821 (American Society of Microbiology, Washington DC, 1996).

Acknowledgements

We thank H. Bujard for the plasmids; J. Rietdorf and R. Pepperkok for help with fluorescence microscopy; M. Diehl and D. Thieffry for discussions; and H. Domingues and R. Guerois for reading the manuscript. A.B. is supported by the Louis-Jeantet foundation.

Correspondence and requests for materials should be addressed to A.B. (e-mail: becskei@embl-heidelberg.de).

Answering life's big questions...

... such as 'how long is a telomere?' or 'found any good genes lately?'.

mVADER

Active Motif www.activemotif.com
mRNA isolation that beats oligo dT

The methodology of mVADER mRNA isolation kits is invasive clamping phosphonopeptide nucleic-acid (pPNA) technology. Poly (A)⁺ RNA is isolated as a triplex with pPNA, formed independently of mRNA secondary structure. The pPNAs used have higher affinity for DNA than for RNA, allowing for differential elution. This, say Active Motif, results in a more representative population of mRNA for arrays or cDNA synthesis than obtained with standard oligo dT methods. mVADER kits come with a proprietary mixture of detergents, proteases and chaotropic salts to ensure denaturation of RNases as well as complete lysis. The kits contain a full set of RNase-free solutions and affinity resins.

Reader Service No. 100

Genscan

eBioinformatics www.ebioinformatics.com
Algorithmic software for BioNavigators

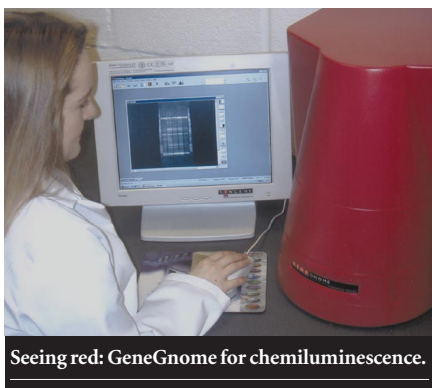
This program uses algorithms derived from mathematical analysis of the relationship between genes and raw DNA sequences to find segments of uncharacterized DNA that are likely to hold undiscovered genes. Genscan picks out potential genes in DNA without the need to work backwards from a known gene product, and is therefore suitable for initiating studies on new data from genome-sequencing laboratories. The program is available for use by subscribers to BioNavigator.

Reader Service No. 101

GeneGnome

Syngene www.syngene.com
CCD camera for chemiluminescence

GeneGnome uses a fixed-focus 16-bit CCD camera in a cabinet designed to allow a short distance between camera and sample. It has a



Seeing red: GeneGnome for chemiluminescence.

wide-aperture lens and is designed to obtain excellent images from any chemiluminescence label, including the low-light-level label ECL. The package includes a flat screen monitor and a keyboard, with GeneSnap and GeneTools software providing automatic image capture, analysis and reporting.

Reader Service No. 102

TOPO cloning kit

Invitrogen www.invitrogen.com
Faster PCR ligation

The pcDNA4HisMax TOPO TA cloning kit combines the functions of the QBI SP163 translational enhancer and TOPO cloning. The enhancer is derived from the 5'-untranslated region of the vascular endothelial growth factor gene and is claimed to increase expression levels two- to fivefold compared with promoter alone. The TOPO cloning kit is provided cloning-ready with 3'-T overhangs that are activated with topoisomerase-I. Invitrogen promise 5-minute bench-top ligation of PCR products by eliminating vector preparation, synthesis of special primers containing restriction sites, modification or purification of PCR products, ligase and overnight incubation.

Reader Service No. 103

Diversify

Clontech www.clontech.com
PCR random mutagenesis kit

The Diversify kit creates a broad range of mutation types, increasing the probability of creating novel functions in altered proteins. Mutation rates of 2 to 8 mutations per 1,000 bp can be obtained by varying the concentration of two buffer components. Additional rounds of PCR yield even higher mutation rates. Diversify uses AdvanTaq Plus polymerase to amplify DNA fragments up to 4.0 kb. More than 1.5 µg of DNA can be obtained from a 50-µl sample even in the most mutagenic buffer conditions, enough DNA for cloning or for constructing a library.

Reader Service No. 104

Barnes maze

Columbus Instruments www.colinst.com
Amaze your friends

Designed for use in neurophysiological and behavioural studies with rats, this stainless-steel maze consists of a circular table, 122 cm in diameter, which can be rotated. Around the circumference are 18 evenly spaced circular 9.5-cm-diameter holes. A fixed box located underneath the platform is constructed



Table 1: exercise your rodents, Barnes maze style.

like a drawer to allow easy removal of the rat. The base of the platform is 140 cm high.

Reader Service No. 105

Immobilon-Ny

Millipore www.millipore.com
Nylon blotting membrane

This membrane is made of uncharged, web-supported nylon 66 with a 0.45-µm pore size and is validated for detecting submicrogram quantities of DNA. Immobilon-Ny is non-sided and non-curling to facilitate colony lifting. Suggested applications include alkaline transfer, radioactive and chemiluminescent detection. The membrane is available in discs, sheets or rolls.

Reader Service No. 106

Telo TAGGG

Roche Diagnostics www.roche.com
Telomere length assay

This assay detects telomeric DNA (TTAGGG) and measures telomere length from cell cultures and other biological samples. After selective enzymatic digestion, purified genomic DNA is separated according to size by gel electrophoresis and transferred to a nylon membrane by Southern blotting. A non-radioactive probe specific for telomeric repeats is hybridized to the DNA fragments and visualized using chemiluminescent substrate.

Reader Service No. 107

Monoclonal antibody service

Biodesign www.biodesign.com
Antibodies on demand

Biodesign's antibody development service offers individual, customized programs for immunization, fusion, subcloning and production. Additional development services include colony expansion, cryopreservation, second facility banking, and ascites production in custom mouse strains or SCID mice. Purification, conjugation and polyclonal antibody development are also available.

Reader Service No. 108

new on the market

Nebula LC/MS

Gilson

www.gilson.com

Mass spectrometry

The Nebula LC/MS purification system is designed for use in drug discovery and biotechnology labs. Nebula analysis systems are paired with ThermoQuest's Finnigan AQA technology to allow operation of the MS for longer periods. Any of the Nebula injectors, from the small-footprint 235 autoinjector to the high-throughput 215 series, can be paired. TopPoint and UniPoint software provide both open-access, multi-user and single-point control, and can perform injection and fraction collection based on UV-Vis detection with subsequent MS fraction analysis for compound qualification and verification or with MS-driven fractionation.

Reader Service No. 109

XL2020-008

Misonix

www.misonix.com

Ultrasonic high-throughput screening

Reactions in 24-, 96- and 384-well microplates can be accelerated using this ultrasonic microplate processor system, and it is particularly recommended for high-



Sounds fast: microplates respond to ultrasound.

throughput screening applications. It is claimed to eliminate cross-contamination by avoiding direct probe intrusion into individual wells. The system comes with microprocessor-based digital circuitry, a Pulsar timer, 600-watt output, variable control, and an LCD screen with operator prompting, elapsed time display and auto overload protection. An optional microplate attachment with stand and bath is available.

Reader Service No. 110

Ettan MALDI-ToF

Amersham Pharmacia Biotech

www.apbiotech.com

High-performance proteomics

This mass spectrometer for protein identification uses Prowl software combined with a control program to match proteins against known databases. It employs proprietary quadratic field reflection technology to perform advanced applications. The spectrometer can be upgraded and is designed so it can be easily integrated with future products planned by the manufacturer.

Reader Service No. 111

Electro-Fast

ABgene

www.abgene.com

Streamlined gel loading

Electro-Fast features an arrangement of combs that form an 8×12 array of wells, with a spacing of 9 mm centre-to-centre, equivalent to standard 96-well microplate format. This enables the use of multichannel pipettes to speed gel loading by a factor of eight or more. The alphanumeric grid location of the samples is preserved, avoiding the need to label multiple tubes. Run times are typically 15 minutes, ideal for rapid screening of PCR products. The Electro-Fast is also available in a 204-well version, and in 'stretch and wide' versions for increased run length, with the distance to each comb doubled to 18 mm.

Reader Service No. 112



On the Prowl: Ettan MALDI-ToF seeks proteins.

BAS-MS 3543

Fujifilm

www.fujifilm.com

A large-format imaging plate

Compatible with most large-format scanners, this imaging plate is said to be 100 times more sensitive than X-ray film. Photo-stimulable phosphor imaging permits the use of an ultra-thin layer to limit radiation scattering. A 14-facet crystal allows denser packing of crystals into each layer. The manufacturer claims that exposure times and radioisotope usage can be reduced by up to 90% even with weaker energy isotopes such as ^{32}P , ^{35}S and ^{14}C . The plate is water-resistant, avoiding the need to freeze or wrap wet samples. Suggested applications include DNA and genetic research, pharmacokinetics, neuroscience and X-ray-diffraction analysis.

Reader Service No. 113

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

mtDNA EXTRACTOR KITS

NEW

- Easy & efficient isolation of mtDNA from whole blood, cell culture or tissue
- Results in ~90 Minutes
- No Hazardous Reagents
- Works even in the presence of anticoagulants

Wako

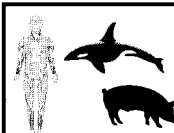
Wako USA 877-714-1921
Wako GmbH (49)02131 3110
www.wako-chem.co.jp

Custom Peptides & Polyclonal Antibodies

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

Research Genetics
U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>



Cross species Primer design with Primer Premier

- ✓ Allele specific primers
- ✓ Degenerate Primers
- ✓ ClustalW alignment
- ✓ Pathogenic primers
- ✓ Restriction enzyme & motif analysis
- ✓ Nested/multiplex primers and more...

Download free demo from
<http://www.PremierBiosoft.com>

Call 888-847-7494 (Toll free in US)
650-856-2703 (International)

PREMIER
Biosoft International
Intuitive software for the molecular biologist

READER ENQUIRY NO. 38

READER ENQUIRY NO. 45

Addressing the US shortfall

More temporary visas could aid US high-tech industry

p597

Germany's IT staffing woes

Chancellor Schroeder takes steps to ease tight labour market

p597

Immigrant impact

Measuring the economic value of highly skilled immigrants

p598

US Congress encouraged to lay out the welcome mat for skilled foreigners

US high-tech industry is crying out for skilled foreign workers, but there just aren't enough visas to go round. Diane Gershon explores...

A critical shortage of skilled labour in the US high-technology industry is forcing Congress to reassess its policy on H-1B temporary visas for well-qualified foreigners. Two years after enacting legislation that almost doubled the annual quota of foreign professional workers entering the United States on such visas, demand is once more outstripping supply.

A number of studies, many by the high-tech industry itself, document the shortages of skilled labour. A survey released last October by the Computing Technology Industry Association found that the shortfall in qualified personnel costs the IT industry \$4.5 billion a year in lost productivity. Almost 10 per cent of jobs (268,740 positions) in IT service and support were unfilled, which translates into a loss of more than \$100 billion for the US economy.

Non-immigrant workers with fairly specialized occupations — those that require a bachelor's degree or higher — can enter the United States with an H-1B temporary visa. This allows them to work in the country for three years, although this can be extended for a further three years. Many H-1B visa holders go on to become permanent residents.

The annual cap on H-1B visas was set at 65,000 in 1990. In fiscal years 1997 and 1998,



Valuable visas: highly skilled foreign workers are helping Silicon Valley to meet its shortfall in staff.

when demand for the visas outstripped supply for the first time, Congress revised the quotas. The cap was set at 115,000 for fiscal years 1999 and 2000, and 107,500 for 2001, as part of the American Competitiveness and Workforce Improvement Act of 1998. The quota is scheduled to drop back to 65,000 in 2002 unless Congress enacts further legislation.

But demand for the visas has never been higher. In 1999, the new cap of 115,000 was reached even earlier than in the previous year

when it was just 65,000. This year, the US Immigration and Naturalization Service announced that demand was up by 50,000 compared with last year, and stopped accepting H-1B visa petitions in March.

A call for action

The National Science Foundation is looking into the shortfall in high-tech workers in a study commissioned by Congress as part of the 1998 Act. But its findings are not expected until after Congress adjourns later this year.

Those who oppose a further increase in H-1B quotas have discredited the available studies. Historically, the US labour unions have viewed the H-1B visa programme as a threat to US workers. They claim it depresses wages and subjects older Americans to age discrimination from companies that prefer to hire younger, cheaper foreign workers.

Nevertheless, the increased demand for H-1B visas, and the fact that any significant shortage of skilled workers could pose a threat to the strength of the US economy, has made Congress sit up and listen. Several members of Congress have introduced bills that would address the issue — at least in the short term.

"We can either allow people to come here on a temporary basis to help expand these companies, or they can very quickly move their operations to where the people are," says an aide to Senator Spence Abraham (Rep-

Germany faces up to its labour shortage

The debate over temporary visas issued to highly skilled workers is not unique to the United States. In Germany, for example, Chancellor Gerhard Schroeder recently proposed a plan to issue 20,000 (10,000 a year for two years) temporary work permits, valid for up to five years, to highly skilled, non-European Union workers to ease the perceived shortage of IT staff.

According to Magnus Lofstrom, a research associate at the Institute for the Study of Labor in Bonn, industry estimates indicate that there are about 75,000 jobs currently available in the IT sector. The temporary immigrants are expected to come mostly from India and Eastern Europe, says Lofstrom and, although the proposal has not yet

been approved, he expects some version of it (with the quota intact) to be passed by the summer.

At 9.8 per cent, Germany has an overall unemployment rate that is more than twice that of the United States. "The higher unemployment rate makes the proposed increase in immigration more vulnerable to criticism," says Lofstrom. "Many Germans feel that it is wrong to look abroad when there are so many unemployed German workers."

The somewhat improved German economy, which has shaved almost one per cent off the jobless total in a year, might make Schroeder's proposal more palatable, says Lofstrom. But the question of whether Germany can entice 20,000 foreign IT workers to its shores is a valid concern, he says.

D. G.

lican, Michigan). Together with Senator Orrin Hatch (Republican, Utah), Abraham has co-sponsored a bill to raise the cap on H-1B visas to 195,000 for 2000, 2001 and 2002. Colleges and universities would be exempt.

In the House of Representatives, David Dreier (Republican, California) and Zoe Lofgren (Democrat, California) have introduced a bill that would up the quota to 200,000 for 2001, 2002 and 2003. But rather than exempting colleges and universities, the bill would set aside a certain number of visas for them, as well as for applicants educated to levels of Master or the equivalent. Both bills contain provisions for educating and training US workers.

The complication in all of this is a bill introduced by Representative Lamar Smith (Republican, Texas), which would put no limits on H-1B visas over the next three years. However, an aide says Smith insists that the bill be balanced and include some provisions that are good for US workers. Compa-

nies would need to show, for example, that the median wage of their US employees had increased over the previous year. There is also a \$40,000 annual minimum wage requirement for H-1B workers, although colleges and universities would be exempted. In addition, the bill contains measures to protect against abuses of the system by so-called job shops set up solely to hire H-1B workers. Moreover, work experience would no longer be an acceptable substitute for a degree.

"Even though on the face of it he [Smith] eliminates the cap altogether, all these other provisions would serve to restrict the usefulness of the programme," says Jeanne Butterfield, executive director of the American Immigration Lawyers Association. "A \$40,000 salary floor ... may sound like a nice round number but it doesn't take into account the particularities of different pockets of the country where that might not be an appropriate equivalent wage." Butterfield says the association supports the Hatch-Abraham and

Dreier-Lofgren initiatives but she feels that Smith's bill is a "political ploy to try and undercut Dreier-Lofgren". As this is a bipartisan effort, Republicans would be unable to claim total credit for it should it pass, she says.

Breaking the bottleneck

The Biotechnology Industry Organization (BIO), a trade association representing 920 of the estimated 1,300 US biotechnology companies, has also thrown its weight behind the Hatch-Abraham and Dreier-Lofgren bills. "The bottlenecks on personnel can be quite damaging," says Chuck Ludlam, BIO's vice-president for government relations. According to the BIO, there are 153,000 people employed in the US biotech industry, about 10 per cent of whom have H-1B visas.

Robert Tjian, a professor of molecular and cell biology at the University of California, Berkeley, understands about the tight labour market. He co-founded the biotech company Tularik in San Francisco about 10 years ago, and says that Tularik "is constantly in the market to hire more chemists".

Although the current policy debate about US immigration is focused almost entirely on highly skilled workers and the H-1B visa quota, there is a larger issue here, says Representative Smith's aide. When Congress began debating the H-1B issue again last year, Smith called it "the tip of the workforce iceberg", referring to the mismatch between the skills of permanent immigrants admitted to the United States and the needs of the workforce. "The reason why we are spending so much time and energy on temporary, basically foreign guest worker visas, is that our legal immigration system is not meeting the nation's needs," says Smith's aide. Today, 35 per cent of the almost one million legal immigrants admitted to the United States each year lack a high-school education, Smith's aide notes. At the same time, 90 per cent of new jobs require more than a high-school diploma.

"Historically in this country in times of economic success ... people are willing to be quite generous in terms of immigration policy: in times of recession and hardship, that opinion turns around," says Butterfield.

Few would argue that more needs to be done in education and training to encourage more Americans to go into computer sciences, engineering and other high-tech disciplines. But the high-tech industry argues that there are job vacancies that need filling today, and this is increasing pressure on Congress to ease restrictions on temporary foreign workers as a stop-gap measure. Although it may help with critical labour shortages in certain high-tech sectors, "it's a short-term fix for a long-term challenge", says Senator Abraham's aide.

♦ <http://www.senate.gov/~abraham>

♦ <http://www.house.gov/dreier>

♦ <http://www.house.gov/lamarsmith>

♦ <http://www.comptia.org>

The economic impact of Silicon Valley's immigrant entrepreneurs

A study called *Silicon Valley's New Immigrant Entrepreneurs* suggests that, rather than simply displacing native-borne workers, as is sometimes argued, many highly skilled immigrants make significant contributions to the economic development of the region.

The study, conducted for the Public Policy Institute of California by AnnaLee Saxenian of the department of city and regional planning at the University of California, Berkeley, showed that foreign-born workers account for about a third of the skilled scientific and engineering workforce in Silicon Valley, with Indians and Chinese predominating. Moreover, Saxenian found that in 1998, Chinese and Indian immigrants

were running a quarter of the high-tech businesses in Silicon Valley, collectively accounting for more than \$16.8 billion in sales and over 58,000 jobs.

The primary motivating force for many of these immigrants to start companies "was the experience of a glass ceiling in established companies", says Saxenian. Gaining access to venture capital has historically been a problem for these groups, she explains. But that's changed a lot, she says, with the formation of a host of highly organized ethnic networks and associations. High-visibility business successes by the likes of Jerry Yang of Yahoo and Sabeer Bhatia of Hotmail have also "alerted the mainstream".

Saxenian says her research underscores important changes in the relationship between immigration, trade and economic development. She now feels that the immigration debate needs to be broadened to reflect the role of highly skilled immigrants in generating jobs, wealth for regional economies, as well as economic links back to their countries of origin.

Diane Gershon is Assistant Editor, New Technology, at *Nature Medicine*.

♦ <http://www.pplic.org>



Rich rewards: Silicon Valley has benefited greatly from immigrant entrepreneurs such as Sabeer Bhatia.